**Invited paper**

Philip Strömert*, Johannes Hunold, André Castro, Steffen Neumann and Oliver Koepler

# Ontologies4Chem: the landscape of ontologies in chemistry

https://doi.org/10.1515/pac-2021-2007

**Abstract:** For a long time, databases such as CAS, Reaxys, PubChem or ChemSpider mostly rely on unique numerical identifiers or chemical structure identifiers like InChI, SMILES or others to link data across heterogeneous data sources. The retrospective processing of information and fragmented data from text publications to maintain these databases is a cumbersome process. Ontologies are a holistic approach to semantically describe data, information and knowledge of a domain. They provide terms, relations and logic to semantically annotate and link data building knowledge graphs. The application of standard taxonomies and vocabularies from the very beginning of data generation and along research workflows in electronic lab notebooks (ELNs), software tools, and their final publication in data repositories create FAIR data straightforwardly. Thus a proper semantic description of an investigation and the why, how, where, when, and by whom data was produced in conjunction with the description and representation of research data is a natural outcome in contrast to the retrospective processing of research publications as we know it. In this work we provide an overview of ontologies in chemistry suitable to represent concepts of research and research data. These ontologies are evaluated against several criteria derived from the FAIR data principles and their possible application in the digitisation of research data management workflows.

**Keywords:** Cheminformatics; FAIR data; linked data; ontology; research data; terminology.

## Introduction

Research data is more than the aggregation of numbers or images in a scientific journal article, experimental section, or supplementary information. To fully reproduce the deduction of the results, we need access to the raw data and how it was generated, processed and analyzed. But simply publishing all this raw data and information somewhere on the web to allegedly make one's research more transparent is not the solution. We need this research data to be FAIR, that is Findable, Accessable, Interoperable and Reusable not only by humans but also machines [1, 2]. While domain experts should, due to their training and implicit knowledge, be able to grasp and interpret the semantics expressed in texts, tables, and images of articles and their experimental sections, computers cannot fully do so without fine grained metadata annotations. Ontologies, taxonomies, terminologies or vocabularies can be used to semantically describe research data, producing this FAIR and machine-readable data. From the perspective of informatics simply put, an ontology is a collection of

**Article note:** A collection of invited papers on Cheminformatics: Data and Standards.

***Corresponding author: Philip Strömert**, TIB – Leibniz Information Centre for Science and Technology, Welfengarten 1 B, 30167 Hannover, Germany, e-mail: philip.stroemert@tib.eu. https://orcid.org/0000-0002-1595-3213
**Johannes Hunold, André Castro and Oliver Koepler,** TIB – Leibniz Information Centre for Science and Technology, Welfengarten 1 B, 30167 Hannover, Germany. https://orcid.org/0000-0002-4378-6061 (J. Hunold). https://orcid.org/0000-0002-7839-3698 (A. Castro). https://orcid.org/0000-0003-3385-4232 (O. Koepler)
**Steffen Neumann,** Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle, Germany. https://orcid.org/0000-0002-7899-7192

machine- and human-interpretable concepts and relations that represent entities and their interdependence in a specific domain. These concepts and relations can be used for the creation of metadata, providing a formalized and in-depth description of research data. The NFDI4Chem project has been created to foster a FAIR research data management (RDM) in chemistry. An important part of the project is the development and improvement of standards for the description of research data via metadata and based on ontologies together with the chemical community, starting with a focus on molecules, their characterisation data and involvement in reactions. Quite a few chemical ontologies already exist to describe the chemistry domain, molecules, their characteristics and reactions (see Table 1 & Fig. 1).

Identifying which of the existing ontologies can be reused for FAIR RDM is the scope of the present paper. The criteria to include an ontology in this review will be elaborated in the methods sections. In the

**Table 1:** Ontologies representing concepts for research data management in chemistry.

| Ontology | Domain | License | Modularity | Used in |
|---|---|---|---|---|
| ChEBI | Chemistry | CC-BY 4.0 | BFO & OBO based | YMDB, HMDB, PubChem, MassBank, KNApSAcK, UM-BBD, GMD, SMID-DB |
| CHIRO | Chemistry | CC0 1.0 | BFO & OBO based | Unknown |
| ChemOnt | Chemistry | Custom OA license | Subsumable under BFO's *Material entity* | YMDB, HMDB, T3DB, ECMDB, DrugBank, PubChem, ChEBI, LIPID MAPS, MoNA |
| CHEMINF | Chemistry | CC-BY 3.0 | BFO & OBO based | PubChem, Open PHACTS |
| CHMO | Chemistry | CC-BY 4.0 | BFO & OBO based | Chemotion, Allotrope™ |
| MOP | Chemistry | CC-BY 4.0 | BFO & OBO based | RXNO |
| RXNO | Chemistry | CC-BY 4.0 | BFO & OBO based | NameRXN, Wikipedia, Chemotion |
| OntoKin | Chemistry | Unknown | OntoCAPE upper level & modules | J-Park Simulator |
| AFO | Chemistry | CC-BY 4.0 | BFO classes & relations, many AFO- some custom OBO-modules | Allotrope™ |
| PROCO | Chemistry | CC-BY 4.0 | AFO & OBO based | Allotrope™ |
| MS | Chemistry | CC-BY 4.0 | BFO & OBO mapping possible | mzML |
| nmrCV | Chemistry | Public Domain Mark 1.0 | BFO & OBO mapping possible | MetaboLights, HMDB |
| BFO | Upper level (classes only) | CC-BY 4.0 | OBO backbone | ¿300 ontologies & ¿50 organizations, PubChem |
| RO | Upper level (relations) | CC0 1.0 | BFO & OBO based | Monarch Initiative, OBO Foundry, Gene Ontology, PubChem |
| IAO | Information artifacts | CC-BY 4.0 | BFO & OBO based | OBO Foundry, Allotrope™, PubChem, ISA tools |
| OBI | Biomedicine | CC-BY 4.0 | BFO & OBO based | OBO Foundry, Allotrope™, PubChem |
| UO | Scientific units | CC-BY 4.0 | BFO & OBO based | OBO Foundry, UOM, PubChem |
| QUDT | Scientific units | CC-BY 4.0 | BFO & OBO based mapping possible | Open PHACTS |
| PATO | Phenotypic & physical qualities | CC-BY 3.0 | BFO & OBO based | OBO Foundry, Allotrope™ |
| SIO | Upper level | CC-BY 4.0 | BFO alignment | PubChem, Bio2RDF, SADI Semantic Web Services, DisGeNET's gene-disease associations, EBI's Gene Expression Atlas, Graph4Code |
| EDAM | Life-sciences & data management | CC-BY 4.0 | BFO & OBO mapping possible | EMBOSS, Bio-jETI |
| OntoCAPE | Upper level & engineering | GNU GPLv2 | Provides upper level concepts | J-Park Simulator |

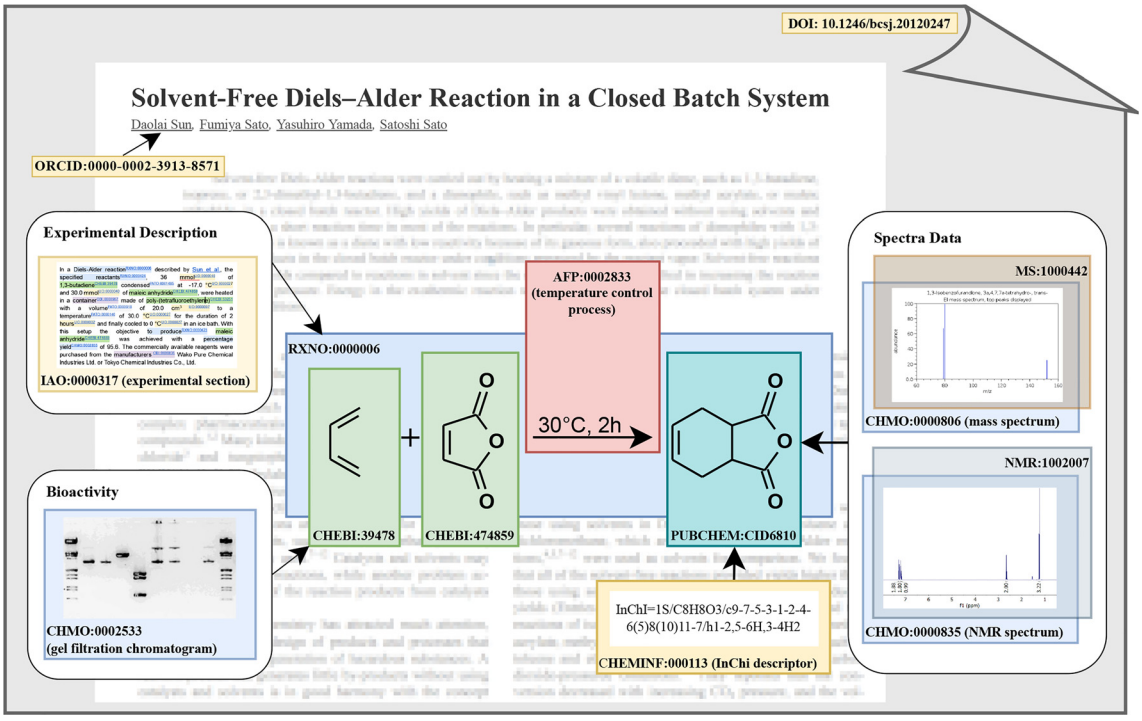


**Fig. 1:** Semantics hidden in a research article on the example of Sun *et al.* [3].

results section the identified ontologies will then be analyzed with regard to the degrees in which they meet these criteria. This includes a brief description of what would have to be done to use an ontology that lacks some of the needed features. Due to the rather broad scope of this paper being an overview, the latter can only be a first starting point and a further in-depth analysis for each identified ontology will have to be carried out elsewhere.

## Background

The true meaning of ontology is tied to the centuries long discourse of philosophers asking the question 'what is' and to the problems that arose from their various answers [4]. Engaging in that discourse would be out of the scope of this paper, and oftentimes confuses the domain users of ontologies. Instead, we will follow a less philosophical but more pragmatic definition of ontology, which has had a significant impact on information science, and which defines ontology as a formally specified conceptualization that is focused on answering the question of what can be represented in a specific domain [5, 6]. Following this definition allows us to interpret the common ways in which chemists have codified their knowledge of molecules, reactions, and the underlying chemical mechanisms using common symbols (*e.g.*, reaction arrows), structural formulas, and systematic names as a kind of ontology. These representations of chemical entities, their properties and relationships are today formalized by the recommendations and terminology collected in the eight IUPAC color books [7]. In the 1980s, John Gorden *et al.* proposed to represent this knowledge in a very chemistry specific formal language based on set theory, first-order logic and the aim to use it in a computational context [8]. The more recent description logic (DL) based syntax and semantics specified by the World Wide Web Consortium, in form of the Resource Description Framework (RDF) [9], and its schema (RDFS) [10] as well as the Web Ontology Language (OWL) [11] facilitates a broader, interdisciplinary exchange of research data over the

internet. Using these specifications, data that is semantically annotated with ontologies as well as ontologies themselves can be stored in the form of 'subject-predicate-object' triples.

In order to speak about ontologies in this overview, we need to introduce some of the most important technical terms. The terms of an ontology used to represent those portions of reality that exist as generic entities, such as atoms and molecules, chemical reactions, lab equipment and experimental methods, are commonly called **classes**. Particular entities that exist are commonly called **instances**, **individuals** or **particulars** of a class. For example, a particular mass spectrometer (*e.g.* identified via a serial number) could be represented in an ontology about lab equipment as an instance of the class *MassSpectrometer*. The terms used to signify the interdependence between classes or individuals are usually referred to as **relations** or **properties**. Depending on the expressiveness of an ontology, properties can be further restricted by being declared transitive, symmetric/asymmetric, functional/inverse functional or reflexive/irreflexive. With domain and range restrictions on a property its applicability can be narrowed down to only those classes or instances thereof between the relation should hold true. Such property restrictions help a reasoner, also called inference engine, to automatically classify given instances to be of a certain kind or to detect logical inconsistencies within an ontology. In OWL properties are further distinguished into **object properties**, **data properties** and **annotation properties**. Whereas object properties exclusively define relations between classes or their instances, data properties are only used to specify relations between classes or their instances and literal values or certain standard XML schema data types such as integers, strings or datetimes. Annotation properties, as the name suggests, are only allowed to be used to provide metadata for the classes, properties, instances as well as the ontology itself (*e.g.* label, definition, comment, creator or examples of use) and are ignored by reasoners.

The backbone of an ontology is the hierarchy of its core classes. This backbone is usually a **taxonomy**, where an *is_a* (in OWL *subclassOf*) relation is used to further distinguish concepts or entities into subclasses (*e.g. MassSpectrometer is_a Device*). Another form of hierarchical structuring in an ontology can be done by grouping classes using the relations *part_of* and *has_part*, specifying a **partonomy**. Similar to the differentiation of classes into subclasses, properties can also be further differentiated into sub-properties. When speaking about a certain part of a taxonomy or a partonomy that includes a root class and its associated classes, the term **branch** is often used.

Another important aspect of representing knowledge in an ontology are the **axioms** postulated in it using description logic. Axioms are the rules defined within an ontology with which to express relations that always hold true between classes or instances of classes. All the *is_a* relations making up the taxonomy of an ontology can thus be understood as the most basic set of axioms. More complex axioms are usually used to logically define a class further and to allow making inferences using a reasoner. An example could be an axiom in a hypothetical ontology that defines the class *MolecularProcess* as a *Process* which must have instances of the class *MolecularEntity* as its participants. The more axioms an ontology contains, the more expressive it is and the stronger is its ontological commitment with regard to the knowledge (conceptualization) it is supposed to represent. By arranging conceptually related classes, relations, and axioms into **modules** or **subsets** of an ontology, the development of large and complex ontologies as well as a partial reuse by external ontologies can be made easier.

With regard to the possible portions of realities covered by ontologies, we also need to elucidate the terms **upper ontology**, **domain ontology** and **application ontology**. Upper ontologies, also known as upper- or top-level and foundational ontologies, cover reality at the most general level, which means their scope is set to formalize the most generic concepts (*e.g.* time, space, objects, processes, qualities or information) and relations (*e.g.* parthood, causality, time and space dependance) [4, 12, 13]. The use of an upper ontology can be challenging outside the ontology development community, as its logical structure and term definitions depend heavily on philosophical knowledge and positions. However, the benefit of using such an ontology becomes evident when classes or modules from several ontologies have to be combined to expand existing ontologies or to describe knowledge spanning different domains [14].

Domain ontologies on the other hand are solely focused to formalize the concepts and relations of a specific domain (*e.g.* chemical substances or chemical processes). What distinguishes domain ontologies from

application ontologies is the fact that the latter can be understood as concrete customized implementations of the former in a certain software application context (*e.g.* the ontology of PubChem) [15].

# Methods

## Screening process

The screening approach used in this overview is an expert-based one that combines the existing knowledge in our project to identify suitable ontologies with a systematic look-up of ontologies indexed by EBI OLS [16], BioPortal [17], and the OBO Library [18]. As an orientation and starting point we also used the overview papers from Batchelor [19] Hastings *et al.* [20] and Gomez-Perez *et al.* [21].

## Selection criteria

To be included in this overview, an ontology had to meet the criteria of being: in scope of a defined set of chemical subdisciplines, made by domain experts, published and maintained in a FAIR way as well as being used in established applications.

With regard to being in scope of the chemical subdisciplines covered by the NFDI4Chem project at the moment, an ontology should provide classes and relations that can be reused to sufficiently and semantically describe research data in organic, inorganic, physical, analytical, macromolecular and pharmaceutical chemistry as well as biochemistry. A dataset is sufficiently described if it can be easily found, accessed and reused in an interoperable way by humans and machines. With regard to the semantic description of research data, the concrete needs and goals may vary, depending on the involved stakeholders. Hence, the ontologies do not necessarily need to use the full expressive power of description logic. Although this would be a major benefit, to be prepared for the future in terms of digitalization of science, we are also interested in finding available chemistry domain-specific taxonomies and other structured controlled vocabularies. Thus, such 'shallow' ontologies are also included. While other chemistry related domains such as engineering or material science are not yet considered in this review, our methodology is readily applicable and ontologies not yet considered in this review might become relevant and can be added to the collection in the future.

Since a domain-specific ontology codifies domain knowledge, its development and maintenance relies heavily on experts who provide this knowledge. Therefore, we were interested in ontologies that have been **developed by** or in collaboration with such **domain experts** and that adhere to best practices in ontology development. This is a mandatory prerequisite to assure a certain level of academic quality and semantic soundness. Such ontologies are preferred not only in terms of their reusability in different use cases, but also because it is likely to be easier to extend and improve them through further expert collaboration.

As previously stated, our major goal is to promote FAIR research data management. By transitivity, any ontology used in this process **must** therefore also **be FAIR** in the meaning of findable, accessable, interoperable and reusable. It must be indexed in prominent registries, such as OLS, Ontobee or BioPortal and must be both technically accessible via a permanent URL and cognitively understandable, *i.e.*, through sufficient documentation. What is deemed as sufficient might vary from the user's perspective. However, the domain and scope, the chosen design patterns, the rationale behind the reuse of terms from other ontologies or the used degree of axiomatisation as well as the competency questions of the ontology in question should definitely be somehow documented. In the best case, this is present in the form of metadata within the ontology itself combined with a paper or manual. The availability of machine-readable license information, to determine if it is allowed to reuse and/or modify an ontology and under what conditions is a mandatory prerequisite for being a FAIR ontology. Preferred are those ontologies that make use of open licenses.

Interoperability difficulties typically arise when ontologies are aligned to different incompatible upper ontologies, to upper ontologies of different versions (*e.g.* BFO 1.1, BFO 2.0 or BFO 2020) or to no upper ontology

at all, as well as when terms of different ontologies share a name and similar definition but are incompatible with regard to their taxonomic position. Data annotated with such similar ontology terms need to be mapped to a common ground in order to be combined and interpreted in a larger context. This greatly hinders the ease of reuse of this data, as well as a rapid prototyping of semantic applications that depend on it. We are thus primarily interested in identifying ontologies that are aligned with compatible upper ontologies or provide a grounding mapping that enables them to be **reusable in a modular way** with automation tools such as ROBOT [22], ODK [23] or OntoFox [24].

Similar to developing software code, ontologies will never be perfect and might contain bugs, lack some needed classes, properties, axioms or metadata and may not be documented well enough for external users. It is therefore also very important that an ontology is being actively and openly maintained and that there are open ways in which bugs, questions or other issues can be filed and discussed with the developers and curators. This is especially important for the expansion of existing domain ontologies. An ontology that is widely **used in prominent applications** is a good indicator for its reusability. A great theoretical ontology that is not embraced in a real-world application is only of limited use to us, as we seek practical solutions to the problems of FAIRly storing research data.

## Results

We have identified 10 ontologies that cover general scientific domains and 12 chemistry domain-specific ontologies (see Table 1). Of these 12, 10 are domain ontologies and two are application ontologies (see Fig. 2).

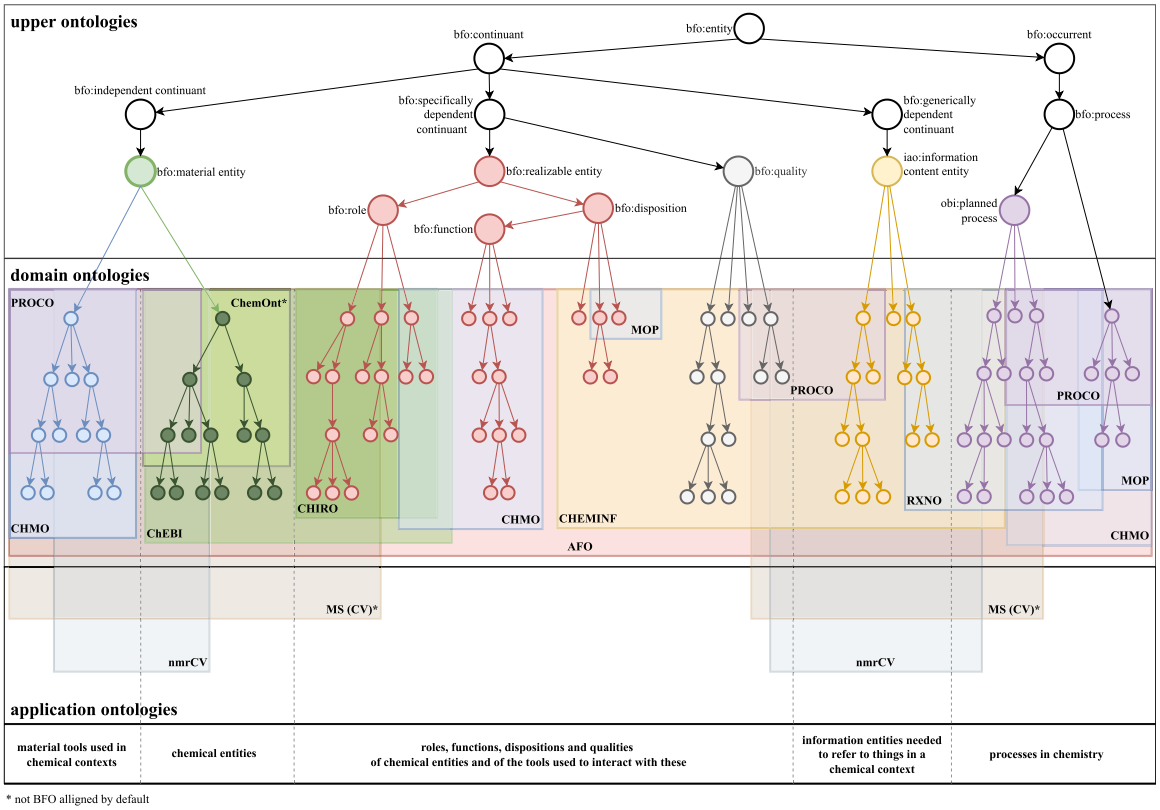


**Fig. 2:** Ontologies relevant for RDM and the scope of NFDI4Chem with regard to their position in the OBO framework.

## General scientific ontologies

Of the 10 general scientific ontologies identified, three, namely BFO, SIO, and OntoCAPE, are upper level ontologies. There are of course many other upper ontologies, like DOLCE, CYC or SUMO [25], however they are not included here due to their limited relevance in the chemical domain. The most important one is the Basic Formal Ontology (BFO), which is used as a unifying reference framework by many renowned ontologies, some of which are also presented later in this review [26]. The BFO development was driven by the need to provide a general framework for the Gene Ontology (GO) [27] and its success led to the formation of The Open Biological and Biomedical Ontology (OBO) Foundry [28, 29]. The community-driven OBO Foundry project has since been known for its library of interoperable ontologies for the life sciences, which are meant to implement the OBO Foundry's best practices (the OBO principles), by using commonly shared design patterns and development tools. The second identified upper-level ontology is the Semanticscience Integrated Ontology (SIO), which, similar to BFO, provides top level classes and relations to describe objects, information and processes as well as their associated basic characteristics (*e.g.* functions, qualities and roles) [30]. According to the developers its design patterns are simple to use in many domains, especially in chemistry, biochemistry, biology and bioinformatics. SIO has been used in projects such as Bio2RDF [31], SADI Semantic Web Services [32], DisGeNET's gene-disease associations [33], NCBI's PubChem RDF [15], EBI's Gene Expression Atlas [34], and the Graph4Code [35] project. To foster semantic interoperability, the SIO developers have provided a mapping to BFO and the Relation Ontology (RO) [36] for important core classes and relations. The third upper level ontology we identified are the parts of the ontology for computer-aided chemical process engineering (OntoCAPE) that define fundamental concepts and relations such as data structures, part-whole relations, processes, material, time and space or SI units [37, 38]. OntoCAPE is designed to be reusable and extendable in many different contexts of computer-aided process engineering, without the need for other ontologies. However, no evidence could be found to indicate that the meta, upper and conceptual layers of OntoCAPE are being used outside of the network in which it is being developed. Similarly, we could not find any references regarding the interoperability between BFO based ontologies and OntoCAPE. Despite these missing references, OntoCAPE is relevant for modeling the domain of process chemistry, for the kinetic chemistry specific ontology OntoKin [39] and is also being used in the NFDI4Cat project (national research data initiative for catalysis) [40].

The remaining general scientific ontologies can be further distinguished with regard to their scope. The Relation Ontology (RO) is the reference ontology of the OBO Foundry for general relations that can be reused in many different contexts. The Information Artifact Ontology (IAO) [41] plays a similar important role in the OBO framework, as it contains all those terms, such as *symbol*, *document*, *data item* or is *quality measurement of*, that are somehow 'about' other entities and that we need to describe information in a machine-readable way. The Ontology of Biomedical Investigations (OBI) [42] contains many of the common scientific terms needed to describe an investigation or experiment, including its protocols and measuring or assay methods and the equipment used in these planned processes. General physical qualities, such as *temperature* or *weight*, are defined in the Phenotype And Trait Ontology (PATO) [43], while the Unit Ontology (UO) [44] contains the terms defining the SI and derived units needed for the proper description of measurements. As all those OBO registered ontologies are supposed to adhere to the OBO principles, they should be interoperable by default with each other. The Ontology of bioscientific data analysis and data management (EDAM) [45] as well as the ontology of Quantities, Units, Dimensions and Types (QUDT) [46] are ontologies outside of the OBO library that are relevant in our context. EDAM covers research areas (topics), types of data, data formats, a categorisation of algorithm functionality and also aspects of biochemistry and analytical chemistry. It is not aligned with any upper ontology, but widely used in the life sciences to annotate tool registries (Bio-jETI) [47] or the ELIXIR training portal (TeSS) [48]. QUDT is developed and published by the non-profit organization and W3C member QUDT.org, with the aim to provide open source industry standard specifications.

## Chemistry domain specific ontologies, taxonomies and schemas

Within the scope of RDM several ontologies exist, which describe experiments, reactions, molecules, analytical methods, devices, algorithms or chemical data formats. In the following we investigate these ontologies with regards to our criteria and their potential application.

### Chemical entities of biological interest ontology (ChEBI)

The Chemical Entities of Biological Interest (ChEBI) ontology from the European Bioinformatics Institute (EMBL-EBI) is probably one of the most widely used ontologies in the chemical domain, as it provides a comprehensive and well-documented classification of chemical entities [49, 50]. The scope of ChEBI can be subdivided into three ontology modules. The first module, which is the branch that is subsumed under the BFO class *material entity*, covers the aforementioned chemical entities. It contains plural and singular terms, where the plural terms refer to groups of chemical compounds. This is consistent with the widely accepted practice in chemical nomenclature, where classes are often named after a particular representative. For instance, phenols (CHEBI:33853) is a class that includes the specific compound, phenol (CHEBI:15882). ChEBI's second ontology module, the branch that is subsumed under the BFO *role* class, covers the roles (*e.g. acid* or *base role*, *catalyst etc.*) chemical entities can have (be a bearer of) when used or studied in a chemical, biological or an application context. The third module covers subatomic particles. The ChEBI ontology serves as the data model for the ChEBI database. The importance of ChEBI to modern chemistry is demonstrated by the many databases it cross-references, such as Human Metabolome database (HMDB), the Golm metabolome database, MassBank, KNApSAcK, UM-BBD, SMID-DB and the Yeast Metabolome database (YMDB) [51], or the many ontologies that reuse terms from it or map to them, such as the Gene Ontology (GO) or the Human Phenotype Ontology (HPO) [52] and many of the chemistry related ontologies covered in this overview. In addition, it is worth noting that the ChEBI Submission Tool provides a rather simple and straightforward way to submit requests for adding terms, which is a good example of open collaboration with the scientific community, allowing domain experts to contribute even without in-depth ontology knowledge. In order to distinguish the preliminary and third party terms from those that are curated by the EBI team, they are categorized into three subsets reflecting that order with the keywords '1-Star', '2-Star', '3-Star'. A very good overview of what ChEBI is capable of and used for is provided by Hastings and Steinbeck [20]. The wide use of ChEBI, its publication under the open CC-BY 4.0 license and the manual curation by domain experts [50] makes it an ideal candidate for reuse in RDM.

### ChEBI integrated role ontology (CHIRO)

The ChEBI Integrated Role Ontology (CHIRO) is a demonstration of how to make the implicit knowledge contained in ChEBI's role branch explicit by axiomatization [53]. It provides links to other OBO ontologies, such as GOPRO, NCBITaxon, HP or DOID through the use of ad-hoc relations, such as *agonist_of* or *inhibitor_of*. The motivation is to establish direct connections between chemical structures such as small molecules or drugs and their effects. CHIRO can thus be used as an ontology module that extends or enhances ChEBI's role branch. While it seems that further development or maintenance of CHIRO is on hold at the moment, the developers of CHIRO have joined forces with a similar project to connect ChEBI's roles to their targets in other controlled vocabularies [52]. In our context, CHIRO should be considered as an important reference point and an opportunity for collaboration once the need to use formalized chemical roles arises.

### ChemOnt ontology

An alternative vocabulary to ChEBI for the classification of chemical compounds is the taxonomy ChemOnt. ChemOnt has originally been developed for ClassyFire [54], which is widely used for automatic classification of especially organic chemical compounds [55]. It was initially released in 2016 and consists of more than 4800

classes, which are named using a consensus-based nomenclature and described based on the characteristic common structural properties of the containing compounds. The impact of ClassyFire's ability to aid chemical data management and to automatically classify new structures in chemistry and biochemistry is illustrated by its implementation in numerous databases such as YMDB, HMDB, T3DB, ECMDB and DrugBank, as well as PubChem, ChEBI, LIPID MAPS, and MoNA – MassBank of North America [54]. Since ChemOnt is a taxonomy, it lacks relations and axioms to further formalize the chemical knowledge it contains. Thus, it only provides the categories (classes) that are needed by ClassyFire. The actual rules/logic behind the application's automated classification is encoded into its software, which makes it more difficult to update when needed. Considering the reusability of ChemOnt it needs to be mentioned that it is neither aligned to BFO like the OBO ontologies, nor to any other upper level ontology. One could subsume ChemOnt's root category *chemical entities* under the BFO *material entity* branch and thus make it reusable in a BFO context. However, there are a few things to consider when doing so. First, there is the obvious naming pattern (*e.g. Halogen oxides*), which violates the idea of only using singular terms for the representation of a universal category and the capitalization violates the OBO Foundry naming conventions. Since this is only a cosmetic problem, it can probably be neglected. What is more challenging when aligning ChemOnt with BFO is its relation to ChEBI. Both are ontologies classifying and thus describing chemical compounds and one would have to make a decision regarding which one to use. As shown by Hastings *et al.* [55], using ClassyFire for automatic classification together with ChEBI is feasible with good results, yet ChemOnt is only partially mapped to ChEBI. For example, a look at ChemOnt shows that most of the subclasses of *Actinide oxoanionic compounds* class are only mapped to ChEBI via their parent, that is to *actinoid molecular entity* (CHEBI:33498). Some, such as *Salt-like carbides* or *Inorganic isocyanides* have no cross reference to ChEBI at all. However, the major difference remains that ChEBI's ontological commitment is to represent chemical compounds, while the ontological commitment of ChemOnt is to represent the chemical compound classes based on structural elements. So the question of how to utilize both ontologies best using a modular approach that is aligned with BFO as upper level ontology needs to be addressed.

### Chemical information ontology (CHEMINF)

The Chemical Information Ontology (CHEMINF) aims to encode the terms, definitions, and logical axioms of chemical information entities [56]. CHEMINF is intended to serve as a single point of truth for the definition and disambiguation of terms and relations used in the domain of cheminformatics. As stated in the CHEMINF documentation, its scope covers chemical graphs and their various encoding formats, the definition of chemical descriptors, such as InChI or SMILES, commonly used software and algorithms, like the *PubChem software library* or *Lipinski rule of five violation calculation algorithm*, as well as format specifications for chemical data, such as the *MOLfile format specification*. CHEMINF also defines needed chemical qualities, such as *molecular structure* or *charge*, as well as dispositions of chemical entities, like *solubility* or *electronegativity*. Being a highly expressive ontology, the provided axioms of CHEMINF further specify the covered entities in a machine-readable way. Explicitly excluded from its scope are the chemical entities defined in ChEBI, aspects of sequence information covered in the Sequence Ontology [57], and further details regarding the defined algorithms or format specifications. Adhering to the OBO Foundry principles, CHEMINF aligns itself with the OBO framework by being an extension of BFO, IAO, OBI and RO. Regarding the alignment of CHEMINF with the upper ontology BFO, it needs to be noted that, at the time of writing, CHEMINF still imports axioms of the outdated BFO version 1.1, which produce logical inconsistencies when reasoning over it. However, this seems to be problematic only when CHEMINF is completely reused and not when only single terms or modules of it are reused elsewhere. Although the documentation of CHEMINF states that it was planned to have the needed qualities of chemical entities defined in PATO, the alignment with PATO did not take place as of yet. When comparing for example the classes defined under *molecular entity quality* in CHEMINF and those defined under *molecular quality* in PATO, it can be seen that certain qualities are still defined in both. Hence, some semantic harmonization is still needed to adhere to the OBO Foundry principle of orthogonality. With regard to the scope of PATO, it remains debatable however, in how far it really is the right place for some of these very domain

specific qualities of chemical entities. Nevertheless, CHEMINF must be considered a required resource whenever there arises the need of describing the various properties of chemical entities, their measurements or predictions, as well as the software and standards used to express these. The ontology is published under a CC BY 3.0 license. Its impact is visible in prominent applications such as, the semantic annotation of PubChem's database, or in the Open PHACTS project.

### Chemical methods ontology, molecular process ontology, and named reactions ontology (CHMO, MOP, RXNO)

The Chemical Methods Ontology (CHMO), the Molecular Process Ontology (MOP) and the Named Reactions Ontology (RXNO) have been developed under the auspices of the Royal Society of Chemistry (RSC) starting around 2008 and were initially created with the aim to enhance semantic publishing in the RSC Project Prospect. The documentation of the three ontologies is mostly limited to the information provided in the respective repositories hosted on GitHub [58]. Hence, the following analysis is mainly derived from examining their OWL representations.

As stated by Batchelor the CHMO is mainly based on the knowledge codified in the IUPAC Orange Book and RSC's Analytical Abstracts [19]. All of its 2939 classes are provided with textual definitions, of which many are derived from the IUPAC Orange book. Focusing on the experimental methods applied in chemistry, most of these classes reside under the OBI *planned process* branch and the ontology can thus be considered to be an extension of OBI. It takes advantage of the fact that OBI has already defined some very important classes for describing scientific experiments, such as *assay* and *material processing* or *device*. Hence, CHMO expands OBI's *assay* branch by defining many subclasses for assay methods used in applied chemistry, such as *spectroscopy*, *thermal analysis* (including *calorimetry*) or *magnetic resonance method*. It also extends OBI's *material processing* and *device* branches by defining many process steps, such as *distillation*, *extraction*, *synthesis method* or *separation method* as well as lab equipment needed in the chemist's daily work. Other material entities needed in a laboratory context such as *buffer solution*, *filter cake* or *chromatographic phase* as well as planned processes and equipment needed for the *waste management*, the *risk management planning process* and the *hazard reduction* are also defined in CHMO. In order to specify the data involved in chemical experimental processes, the ontology defines terms that fall under the IAO *data item* or the IAO *directive information entities* branch. For the formal logical definitions of certain experimental methods via axioms, CHMO reuses some classes defined in ChEBI's *molecular entity* branch, such as *chloroform* for the *polarimetry of sample dissolved in chloroform*, *polymer* for the *polymer preparation method* or *carbon-13 atom* for the *13C nuclear magnetic resonance spectroscopy*. Looking at the relations between the classes defined in CHMO, it becomes clear that only a few external relations are reused. Five domain-specific relations are defined in CHMO: *has_analyte*, *has_matrix*, *probes_atom*, *prevents* and *mitigates*. With regard to the relations between the devices, the methods and their inputs and outputs, the OBI relations *has specified input*, *has specified output* and their inverse relations are reused. Regarding our selection criteria, CHMO seems to be a very good candidate for reuse. Although semantic features of the Prospect Project are no longer publicly available, we assume that CHMO is still used in the backend of RSC's publishing service as well as ChemSpider. CHMO is also used in the Golm Metabolome Database, the MetaboLights Database and the Chemotion ELN and Repository [19, 59, 60]. There are some minor issues that should be addressed to improve the interoperability. For example most of the external relations reused in CHMO are defined in a BFO version that includes very general and temporalized relations. This violates the OBO Foundry principle 7, which states that RO should be used for general relations whenever possible. Other issues that need to be investigated further elsewhere are the semantic overlap with other ontologies and several decisions regarding the subsumption of some CHMO classes. OBI, for example, has defined quite a few of the devices needed in an NMR experiment, but those are not reused in CHMO, although the OBI class *X-ray source* is expanded by CHMO. Another example would be the CHMO class *concentration* which has no reference to PATOs *concentration*, although these two seem to be referring to the same thing.

The Molecular Process Ontology (MOP) is focused on the definition of general molecular processes, such as *addition reaction*, *cyclization* or *polymerisation*. It is thus a rather small ontology that mainly serves as a basis module for the Named Reaction Ontology (RXNO), in which these fundamental molecular processes are needed for the definition of the more complex reactions. Similar to CHMO, the definitions and synonyms defined in MOP are mostly derived from and linked to their respective IUPAC Gold Book entries. In MOP itself, only the relation *is_catalysis_of* is defined and only used to formally define the class *catalysis* as being a molecular process that *is_catalysis_of* some other molecular process. Also similar to CHMO, chemical entities from ChEBI are reused for the axiomatization of certain molecular processes.

The Named Reaction Ontology (RXNO) expands the molecular processes defined in MOP to cover synthetic organic reactions with small-molecules of which it currently contains 647, such as the class of the well-known Diels-Alder cyclization. The top level classification of RXNO contains reactions that change the skeleton (*e.g.* cleaving, condensation, rearrangement), as well as reactions that preserve the skeleton (*e.g.* addition, elimination, protection or deprotection) [61]. As described by Colin Batchelor the classification of named reactions in RXNO is based on two principles, first comparing the longest carbon chains in the reactants and products and then checking whether a ring system is created, broken or altered [19]. More context is also often provided by linking class definitions to publications and similar to CHMO and MOP, the GO annotation properties for exact, narrow or related synonyms are used. For the formalization of its classification, RXNO also reuses chemical entities from ChEBI as well as classes from OBI and IAO. Many of the named reactions in RXNO are subsumed under the class *planned reaction step* which is a subclass of OBI's *planned process*. Reusing an ontology design pattern from OBI, this also entails defining subclasses of the IAO *objective specification* in order to provide the objectives according to which a planned reaction is supposed to take place. Together with the OBI relation *achieves planned objective* these reaction objective subclasses are used in RXNO to formalize the definitions and categorization of most planned reactions steps and synthesizes it provides. To further formalize these classes, RXNO also provides the six domain specific relations: *protects*, *deprotects*, *has specified product*, *has specified reactant*, *has_catalyst*, and *has_intermediate*. Besides the aforementioned use in RSC's semantification of published papers via text mining RXNO is implemented in the Wikipedia info boxes, the NameRXN (NextMove Software) for automatic classification of reactions with SMIRKS [62], and the Chemotion ELN and Repository for the aid of manual reaction classification [60]. Due to this, its covered domain, its references to other data sources, and the synonyms it provides, RXNO must be considered an important resource for describing the provenance of research data. The use of OBO unsupported BFO relations is, as in the other two RSC ontologies, also in this case a minor issue. Another rather easily resolvable issue concerns the current chosen import strategy of MOP classes, as it leaves room for curation errors. For example, the class *cycloaddition* is subsumed under *cyclisation* in RXNO, but not so in MOP. With a direct import of MOP in RXNO such asynchronous errors could be avoided, as the MOP file would be the single point of truth. Overall all three RSC ontologies pass our criteria in major points and are classified to be relevant in our context.

#### Ontology for chemical kinetic reaction mechanisms (OntoKin)

Being developed and actively maintained by domain experts the Ontology for Chemical Kinetic Reaction Mechanisms (OntoKin) is intended to be used for the simulation and understanding of the behavior of chemical processes and can be seen as a domain specific extension of the OntoCape ontology [39]. With 57 classes, 36 object and 99 data properties and its DL based axioms, OntoKin is a very expressive ontology. Its scope can be divided into five modules: reaction mechanism, phase, chemical reaction, rate coefficient and chemical species. These modules define the classes (*e.g.*, *BulkPhase* or *GasPhaseReaction*), relations (*e.g.*, *hasElement* or *belongsToPhase*), and axioms (*e.g.*, *ChemicalReaction* always has a *Reactant*, *Product*, *ReactionMetadata*, *ReactionOrder* and *StoichiometricCoefficient*) that are needed for a comprehensive semantic description of reaction kinetics. From OntoCAPE the classes *ChemicalReaction*, *ChemicalSpecies*, *ReactionRateCoefficient*, *ThermoModel* and *StoichiometricCoefficient* are reused, but only very few terms from other preexisting ontologies. Interestingly, these classes are not imported together with the object or data properties in which they are defined as domain or range. Although excluding other ontologies apart from OntoCAPE is quite reasonable due

to the different scopes, no links to major chemistry ontologies or databases such as ChEBI or CHEMINF are provided. Thus, OntoKin on the one hand has the advantage of having very few dependencies on other ontologies, which makes it robust in terms of semantic stability and easier to implement in applications, such as the prototype of an open access knowledge base (KB) containing chemical kinetic reaction mechanisms to categorize, relate and validate the empirical data encoded in the CHEMKIN file format [39]. On the other hand however, having such few dependencies also means that data described with OntoKin is not as interoperable as data described with an OBO compliant ontology. OntoCAPE, keeps the annotation of its classes and relations mostly limited to a human readable definition using the 'rdfs:comment' property. In OntoKin this is similar, but in addition the definition source of each term is also provided as an IRI to the OWL source file. This lack of meta-information on the ontology terms itself can be considered a possible hindrance for future users.

#### Allotrope foundation ontology (AFO)

The Allotrope Foundation Ontology suite (AFO), first published in March 2018, is a collection of taxonomies and ontologies developed by the Allotrope Foundation that is intended as a standard language for describing equipment, processes, materials, and results. AFO provides the semantic context in a technology stack, called Allotrope Framework, which also consists of the Allotrope Data Model (ADM) and the Allotrope Data Format (ADF). The goal of ADF is to unify the laboratory IT landscape by becoming the gold standard with regard to the many different data formats present today. According to Millecam *et al.* [63], its adoption is still at the very beginning.

A modular reuse of Allotrope's ontology suite along with OBO based ontologies might lead to problems. AFO is BFO-based, nevertheless there are certain aspects that distinguish it from OBO-based ontologies. One of the most prominent differences might be the decision to strictly follow the principle of Single Inheritance [64], which means that a class in AFO cannot have multiple parent classes. Another noteworthy difference concerns the not so well documented import strategy of BFO as well as classes and relations from other OBO ontologies. The AFO modules are developed according to the Allotrope Framework Term Curation Guide and the modules consisting of classes and relations from external ontologies are manually curated [65]. AFO reuses the BFO 2020 ISO version for the top level classes and relations, whereas the BFO 2.0 version used by the OBO community deliberately leaves out any relations. From IAO the main class *information content entity* is imported with some of its subclasses. However, these subclasses have been subsumed under new parent classes introduced by AFO, namely either *facet*, *proposition*, *registry* or *representation form*. Many of the qualities defined in PATO are reused and extended in AFO. From ChEBI only the classes *molecular entity*, *molecule* and *subatomic particle* among only a small selection of their subclasses are reused. Similarly, only the classes *organism*, *manufacturer* & *manufacturing*, *plan* & *planned process*, *processed material* & *material processing* are reused from OBI. Looking at the relations defined in AFO, it becomes clear that most of the relations defined in RO are reused, but also many new relations are being introduced. Quite a few of the classes defined in AFO seem to already have an OBO equivalent. Yet, such classes seem to be only slightly modified OBO classes, where sometimes the labels and sometimes the textual as well as logical definition have been changed and axioms have been added or left out, see for example the AFO classes *specification*, *action specification* and *plan specification*. This indicates that Allotrope's curation of external ontology modules must somehow also entail the adaptation of external classes and properties into AFO. Unfortunately, this opens up a semantic gap between the preexisting OBO work and AFO, which would have to be addressed, if AFO is to be used in a modular way with OBO ontologies. This lack of semantic harmonization with other OBO work can be considered an example of the differences in the development behind corporate doors and the open source community.

AFO's advantages definitely lie in the fact that it is so tightly integrated into the Allotrope Framework, which also consists of customized tools that enable data validation via SHAQL constraints. With regard to the industry-led vision behind ADF, as well as the resources and stakeholders involved in the development and implementation of the Allotrope Framework, it is safe to say that any semantic description of chemical data must probably somehow entail either a reuse of or a mapping to terms defined in the AFO in order to be

interoperable with industry standards in the future. However, overcoming the differences between AFO and other BFO based ontologies, especially OBO ontologies, will be a major challenge that would best be addressed by a closer collaboration between the OBO community and Allotrope.

#### Process chemistry ontology (PROCO, former OPC)

The Process Chemistry Ontology (PROCO) describes the domain of process chemistry from *route scouting*, *process optimization*, *process validation* and *process maintenance* with key concepts like product quality, production processes, environmental sustainability, regulatory compliance, and safety [66]. This BFO aligned ontology is currently being developed as a joint venture between academia (University of Michigan) and industry (Merck, GSK, Allotrope Foundation). In order to foster the collaboration between the Allotrope Foundation and the OBO community, PROCO was submitted for review to the OBO foundry in April 2021. Following a button-up approach in the definition of certain terms, the PROCO developers identified standardized lab practices and basic process patterns that are explicitly or implicitly present in ADM as well as in the specifications of regulatory agencies. PROCO reuses terms from chemical ontologies, such as CHMO and ChEBI, alongside other needed common OBO and non OBO terms (mainly from AFO and SIO). A look at the alignment with BFO shows with regard to the *material entity* branch, that PROCO reuses ChEBI's *chemical entity* class along with the core subclases *atom*, *group* and *molecular entity*, but also adds process chemistry related subclasses to it like *chemical impurity*, *chemical product*, *crystalline solid*, and *starting material for synthesis*. From CHMO the class *portion of material* is reused as a parent for the AFO classes *portion of mixed material* and *chemical substance*, as well as to introduce classes like *product stream* and *waste stream*. ChEBI's *role* branch is also imported for reuse, but only the *chemical role* class is extended and subsumed under BFO's *role* class. From AFO many more chemical roles as well as functions and qualities are imported. Switching the perspective to the needed standard equipment in the covered chemical processes, it can be seen that PROCO defines classes, like *beaker*, *crucible*, *flask* or *bunsen burner*, and subsumes them under the AFO class *device* instead of the imported OBI class *device*, where OBI's *pipette* or CHMO's *chromatography column* could be found. This is most likely due to the needed alignment with AFO for being able to process data shaped according to ADM. For the required information artifacts, PROCO extends IAO by defining process chemistry relevant data items like *crystallization yield*, *fate of impurity*, *material costs* or *purge factor*, documents like *batch manufacturing record* or *risk assessment*, document parts like *clinical trial application (CTA) sections* by the European Medicine Agency as well as directive information entities like the *ICH Guideline*. As for the processes covered in PROCO, we can note that it mainly reuses and extends OBI's *planned process* branch. Many of CHMO's *material processing* classes are reused to contain PROCO's core classes *batch campaign*, *unit operation in chemical processing*, *process monitoring*, *process safety*, *process validation* and *process chemistry filing*. Making good progress in the development, some work is left before PROCO can be easily reused alongside with other OBO-compliant ones. The review by the OBO community is ongoing and the refactoring from OPC to PROCO has not been started yet. However, due to its scope, the fact that this is a cooperation between industry and academia, and the intention to integrate PROCO as a module in the Allotrope framework with the aim to improve industrial production, means that future developments should be monitored. Also being a 'community-based ontology', makes PROCO a great candidate for reuse.

## Chemistry specific application ontologies

#### Controlled vocabularies by the HUPO-PSI

Under the umbrella of the Human Proteome Organization Proteomics Standards Initiative (HUPO-PSI), many domain experts have formed working groups to develop and actively maintain the Mass spectrometry ontology (MS), the Protein modification ontology (MOD) and the Molecular Interactions Controlled Vocabulary (MI), according to predefined rigorous guidelines that adhere to the OBO principles and other best practices [67].

These application ontologies are designed to serve as structured controlled vocabularies (CVs) that list the standardized terms used as allowed values in the XML-based file format standards developed by the proteomics community, instead of representing the data as instances of classes stored in form of serialized triples. Because data validation is also performed using an XML-based approach, there is no need for a high degree of axiomatization, an alignment with an upper ontology, or many object and data properties within the CVs [68].

Due to the scope of MOD and MI being more in the domain of biology than chemistry, only MS is considered relevant in our current context. Its scope is the description of mass spectrometer output files and the interpretation of mass spectra [69], with the two most important branches anchored in the root classes *spectrum generation information* and *spectrum interpretation*. The other 10 branches contain the classes needed to represent related concepts (*e.g. molecular entity*, *software* or *regular expression*). They are either differentiated into finer subclasses (*e.g. atom*, *Brucker software* or *Cleavage agent regular expression*), or associated with conceptually similar classes using the MS *part_of* relation. For example, the *molecular entity attribute* class, which is defined as part of the *molecular entity* class, provides in its subclasses the attributes needed to further describe a molecular entity (*e.g.* SMILES formula). This partonomy structure provides a grouping of concepts that is directly usable in the software that generates the XML based files that encode MS experiment information. For the representation of other common qualities needed in this domain, MS directly imports PATO, and for SI units as well as common quantity prefixes it directly imports UO. Undocumented unfortunately are the reasons why semantically similar classes are not reused from or aligned with other external ontologies (*e.g.* SMILES formula or InChIKey in CHEMINF, or all the same molecular entities in ChEBI).

#### Nuclear magnetic resonance CV (nmrCV)

The scope of the nuclear magnetic resonance CV (nmrCV) is the conceptualisation of terms needed in the description of nuclear magnetic resonance (NMR) assays. It is developed by experts from the Metabolomics Standards Initiative (MSI) under the governance of the COSMOS EU and PhenoMeNal EU projects and maintained on GitHub. Being designed as a simple taxonomy, nmrCV has no object or data properties. Its primary application context, similar to the PSI CVs, is the creation and validation of XML files storing NMR assay data in the nmrML file format [70]. With regard to its modularity, it can be noted that nmrCV is BFO-based, but otherwise does not really follow the OBO core principle of reusing as much as possible from existing OBO ontologies. For example, such general classes as *software* and *instrument* are defined in nmrCV instead of being reused from IAO or OBI. The same is true for some of the domain specific concepts such as NMR pulse sequences or NMR instruments, for which there are already equivalents in CHMO. A mapping to such close or exact matches in other ontologies is also not provided. This might be the only reason hindering a direct reuse of nmrCV in different application contexts. A semantic harmonization with the existing OBO ontologies, especially with CHMO, is thus a needed improvement to enhance interoperability.

## Discussion

As shown in Table 1, the analysis and comparison of existing chemistry ontologies to our set of criteria, derived from the FAIR principles, provides a first assessment to judge their potential reusability in the context of research data management and the NFDI4Chem project. The previously stated criteria findability and accessibility are provided for all identified ontologies. They can be found in registries such as OLS, Ontobee or BioPortal and can be downloaded from these sources. In the context of accessibility and reusability the availability of license information is mandatory. While most of the ontologies are published under CC-BY 4.0 or a similar open license we were unable to identify license information about OntoKin. The related OntoCAPE has been released under a GNU General Public License. We have determined the reusability and modularity based on an existing or possible alignment to an upper level ontology. Most of the discussed ontologies can be labeled as OBO-based and are aligned to BFO. OntoCAPE on the other side defines fundamental concepts and relations in meta, upper and conceptual, layers thus having characteristics of an upper-level ontology in itself.

OntoKin is aligned to OntoCAPE rather than BFO. Although AFO can be seen as a BFO-based ontology, there are several major issues, like its import strategy of external classes and properties or its strict single inheritance design principle, which make the modular reuse of the Allotrope ontology suite challenging and demanding. AFO's advantage is the integration into the whole Allotrope Framework providing customized tools for data validation. ChemOnt is not directly aligned to BFO, nor to any other upper level ontology. While there is a possible approach to achieve an alignment, this needs further investigation, especially in context with ChEBI. Nevertheless both ChEBI and ChemOnt are most useful to classify and describe molecules. CHEMINF has some minor issues with the alignment to BFO and semantic harmonization, but can be considered a valid resource to describe various properties of chemical entities like InChI or SMILES, their measurements or predictions. CHMO shows some minor issues with BFO alignment as well. We defined the degree of quality and semantic soundness by the reputation of the creators and curators as well as the reuse of the ontology in other ontologies or projects. This can be confirmed for all examined ontologies which are either curated by respected research groups, institutions or learned societies. To name only a few prominent examples of reuse we can state that ChEBI, ChemOnt, and CHEMINF are reused in PubChem's database, CHMO and PROCO are reused in the Allotrope Framework and RXNO in chemotion ELN and NameRXN. Although we could not find a proven application of MOP we approve it with regard to its foundational scope and its role as the core module for RXNO to be a relevant ontology in our scope.

Considering all criteria, perspectives and possible improvements we consider the discussed ontologies suitable for being used in the context of research data management and the NFDI4Chem project. For OntoKin license information needs to be identified. Possible implementations of AFO need to be further examined. Apart from some of the issues mentioned here that need to be addressed in the source files of the ontologies, there are also issues concerning their documentation. As we have seen the documentation of the ontologies varies considerably and is often spread out in multiple academic articles. In several cases even persistent citable references were unavailable. In order to allow them to be used by domain experts and ontology engineers who want to easily contribute to their maintenance and further development, this should be improved and standardized.

## Conclusion

Linking ontologies of multiple disciplines and aligning them to common upper-level ontologies enables the creation of huge semantic dataspaces, an interdisciplinary network of interconnected knowledge graphs. With the identified ontologies we have an initial terminology at hand to start building knowledge graphs based on annotated research data. NFDI4Chem aims to support the RDM in daily lab work, addressing digital workflows with data from planning and conducting experiments, describing reactions and molecules, and various analytical data, as well as their archiving and publication. The presented upper-level, general, chemistry and application ontologies need to intertwine to foster data annotation and management along these workflows supported by tools like ROBOT and ontology-ready ELNs to parse or even create annotated data from the very beginning. Similar to the object-oriented programming (OOP) metaphor from software development, we advise the concept of modularity, import and reuse of terms from existing ontologies. For building a semantic chemical dataspace, we need semantic harmonization. We need to be sure that we are talking about the same things when searching for data. This important insight is also relevant for nonchemists, as the semantically described chemical entities can be related to different contexts. In the end we will be able to provide SPARQL endpoints or similar pathways to a knowledge graph to the community. To reach these goals, we need community-based efforts including both domain and ontology experts. We propose to apply best practices of open source software development. This includes proper documentation of an ontology in terms of its scope, competency questions, used design patterns, naming conventions and use cases (including their implementation in existing projects), as well as openly available and easily accessible source codes in a repository. Luckily, most of the ontologies identified in this paper are maintained openly on GitHub. Only in this way will we be able to develop open and FAIR ontologies.

## Outlook

NFDI4Chem is an essential part of a larger consortium network, which was tasked by the German Research Foundation (DFG) to establish a national research data infrastructure. Some initial starting points for chemistry were outlined in this paper. We will further provide and develop intuitive, easy-to-use, and well documented tools and services aligned with the interests and needs of our community to create, curate, provide and archive ontologies. These tools will be integrated and linked with the NFDI4Chem Terminology Service (TS) as a central platform not only to access ontologies but also to support the curation (*e.g.* term request) and development of ontologies. A first step will be an overarching, less technical view that connects to upstream ontologies using templates and existing APIs of platforms such as GitHub in order to automatically file and track issues of the various ontology repositories. This way our interface will make it easier to track term discussions spanning over multiple ontologies and therefore allow identification of similarities and connections. We further aim to initiate a community process to foster and harmonize ontology development by organizing a series of Ontologies4Chem workshops.

**Research funding:** The presented work was conducted as part of the NFDI4Chem project (DFG project no. 441958208). The authors would like to thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for funding and support.

## References

[1] M. D. Wilkinson, M. Dumontier, I. J. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons. *Sci. Data* **3**, 160018 (2016).

[2] A. Jacobsen, R. de Miranda Azevedo, N. Juty, D. Batista, S. Coles, R. Cornet, M. Courtot, M. Crosas, M. Dumontier, C. T. Evelo, C. Goble, G. Guizzardi, K. K. Hansen, A. Hasnain, K. Hettne, J. Heringa, R. W. W. Hooft, M. Imming, K. G. Jeffery, R. Kaliyaperumal, M. G. Kersloot, C. R. Kirkpatrick, T. Kuhn, I. Labastida, B. Magagna, P. McQuilton, N. Meyers, A. Montesanti, M. van Reisen, P. Rocca-Serra, R. Pergl, S.-A. Sansone, L. O. B. da Silva Santos, J. Schneider, G. Strawn, M. Thompson, A. Waagmeester, T. Weigel, M. D. Wilkinson, E. L. Willighagen, P. Wittenburg, M. Roos, B. Mons, E. Schultes. *Data Intell.* **2**, 10 (2020).

[3] D. Sun, F. Sato, Y. Yamada, S. Sato. *Bull. Chem. Soc. Jpn.* **86**, 276 (2013).

[4] B. Smith. Ontology, in *The Blackwell Guide to the Philosophy of Computing and Information*, L. Floridi (Ed.), Wiley-Blackwell, Malden, USA (2004), https://doi.org/10.1002/9780470757017.ch11.

[5] T. R. Gruber. *Int. J. Hum. Comput. Stud.* **43**, 907 (1995).

[6] A. Rector, S. Schulz, J. M. Rodrigues, C. G. Chute, H. Solbrig. *J. Biomed. Inf.* **100S**, 100002 (2019).

[7] G. J. Leigh. *Principles of Chemical Nomenclature: A Guide to IUPAC Recommendations*, Royal Society of Chemistry, Cambridge (2011).

[8] J. E. Gordon. *J. Chem. Inf. Model.* **28**, 100 (1988).

[9] G. Schreiber, Y. Raimond. *RDF 1.1 Primer*; W3C Note; W3C (2014), https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/.

[10] D. Brickley, R. Guha. *RDF Schema 1.1*; W3C Recommendation; W3C (2014), https://www.w3.org/TR/2014/REC-rdf-schema-20140225/.

[11] *OWL 2 Web Ontology Language Document Overview (Second Edition)*; W3C Recommendation; W3C (2012), https://www.w3.org/TR/2012/REC-owl2-overview-20121211/.

[12] G. Amaral, F. Baião, G. Guizzardi. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **11**, e1408 (2021).

[13] C. Maria Keet. in *The Semantic Web: Research and Applications*, pp. 321–335, Springer, Berlin, Germany (2011).

[14] L. Temal, A. Rosier, O. Dameron, A. Burgun. in *Studies in health technology and informatics*, pp. 1065–1069, IOS Press, Amsterdam, Netherlands, 160th ed. (2010), https://ebooks.iospress.nl/publication/13605 (accessed Mar 15, 2022).

[15] G. Fu, C. Batchelor, M. Dumontier, J. Hastings, E. Willighagen, E. Bolton. *J. Cheminf.* **7**, 34 (2015).

[16] S. Jupp, T. Burdett, C. Leroy, H. Parkinson. in *Proceedings of the 8th International Conference on Semantic Web Applications and Tools for Life Sciences*, Vol. 1546, pp. 118–119, CEUR-WS, Aachen, Germany (2015).
[17] N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. L. Rubin, M.-A. Storey, C. G. Chute, M. A. Musen. *Nucleic Acids Res.* **37**, W170 (2009).
[18] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, The OBI Consortium, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, S. Lewis. *Nat. Biotechnol.* **25**, 1251 (2007).
[19] C. Batchelor. in *The Future of the History of Chemical Information; ACS Symposium Series*, 1164, pp. 219–235, American Chemical Society, Washington, DC, USA (2014).
[20] J. Hastings, C. Steinbeck. in *Handbook of Computational Chemistry*, J. Leszczynski, A. Kaczmarek-Kedziera, T. Puzyn, M. G. Papadopoulos, H. Reis, M. K. Shukla (Eds.), pp. 2163–2181, Springer International Publishing, Cham (2017).
[21] A. Gómez-Pérez, M. Martínez-Romero, A. Rodríguez-González, G. Vázquez, J. M. Vázquez-Naya. *Curr. Top. Med. Chem.* **13**, 576 (2013).590.
[22] R. C. Jackson, J. P. Balhoff, E. Douglass, N. L. Harris, C. J. Mungall, J. A. Overton. *BMC Bioinf.* **20**, 407 (2019).
[23] N. Matentzoglu, C. Mungall, D. *Goutte-Gattat. Ontology Development Kit (1.2.29)*, Zenodo, Genève, Switzerland (2021). https://doi.org/10.5281/zenodo.5788237.
[24] Z. Xiang, M. Courtot, R. R. Brinkman, A. Ruttenberg, Y. He. *BMC Res. Notes* **3**, 175 (2010).
[25] V. Mascardi, V. Cordi, P. Rosso. in *WOA 2007: Dagli Oggetti agli Agenti. 8th AI*IA/TABOO Joint Workshop "From Objects to Agents": Agents and Industry: Technological Applications of Software Agents, 24-25 September 2007, Genova, Italy*, pp. 55–64, Seneca Edizioni, Torino, Italy (2007), http://woa07.disi.unige.it/papers/mascardi.pdf (accessed Mar 15, 2022).
[26] R. Arp, B. Smith, D. A. Spear. *Building Ontologies with Basic Formal Ontology*, MIT Press, Cambride, London (2015).
[27] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock. *Nat. Genet.* **25**, 25 (2000).
[28] K. Munn, B. Smith. in *Applied Ontology: An Introduction*, De Gruyter, Boston, USA (2013).
[29] O. Foundry. The OBO Foundry, http://www.obofoundry.org/ (accessed Sept 23, 2021).
[30] M. Dumontier, C. J. Baker, J. Baran, A. Callahan, L. Chepelev, J. Cruz-Toledo, N. R. Del Rio, G. Duck, L. I. Furlong, N. Keath, D. Klassen, J. P. McCusker, N. Queralt-Rosinach, M. Samwald, N. Villanueva-Rosales, M. D. Wilkinson, R. Hoehndorf. *J. Biomed. Semant.* **5**, 14 (2014).
[31] A. Callahan, J. Cruz-Toledo, M. Dumontier. *J. Biomed. Semant.* **4**, S1 (2013).
[32] A. Riazanov, J. B. Laurila, C. J. O. Baker. *BMC Bioinf.* **12**, S6 (2011).
[33] N. Queralt-Rosinach, J. Piñero, À. Bravo, F. Sanz, L. I. Furlong. *Bioinformatics* **32**, 2236 (2016).
[34] EMBL-EBI RDF platform Linked Open Data platform for EBI data, https://www.ebi.ac.uk/rdf/documentation/expressionatlas/ (accessed Sept 23, 2021).
[35] I. Abdelaziz, J. Dolby, J. P. McCusker, K. Srinivas. Graph4Code: A Machine Interpretable Knowledge Graph for Code, arXiv: 2002.09440 (2020).
[36] C. Mungall, N. Matentzoglu, D. Osumi-Sutherland, pgaudet, Clare72, J. A. Overton, J. Balhoff, S. Moxon, N. Harris, M. Brush, V. Touré, B. Duncan, M. Sinclair, sabrinatoro, J. Poelen, A. Bretaudeau, S. Cain, M. Haendel, N. Vasilevsky, diatomsRcool, J. Hammock, M.-A. Laporte, M. Jensen, M. Larralde. oborel/obo-relations: Release 2021-08-31 (2021), https://doi.org/10.5281/zenodo.5347723.
[37] J. Morbach, A. Yang, W. Marquardt. *Eng. Appl. Artif. Intell.* **20**, 147 (2007).
[38] W. Marquardt, J. Morbach, A. Wiesner, A. Yang. in *OntoCAPE: A Re-Useable Ontology for Chemical Process Engineering*, W. Marquardt, J. Morbach, A. Wiesner, A. Yang (Eds.), pp. 353–368, Springer Berlin Heidelberg, Berlin, Heidelberg (2010).
[39] F. Farazi, J. Akroyd, S. Mosbach, P. Buerger, D. Nurkowski, M. Salamanca, M. Kraft. *J. Chem. Inf. Model.* **60**, 108 (2020).
[40] C. Wulf, M. Beller, T. Boenisch, O. Deutschmann, S. Hanf, N. Kockmann, R. Kraehnert, M. Oezaslan, S. Palkovits, S. Schimmler, S. A. Schunk, K. Wagemann, D. Linke. *ChemCatChem* **13**, 3223 (2021).
[41] A. Ruttenberg, A. Goldstein, A. Goldfain, B. Smith, B. Peters, C. Torniai, C. Mungall, C. Stoeckert, C. A. Boelling, D. Natale, D. Osumi-Sutherland, G. Frishkoff, H. Stenzhorn, J. A. Overton, J. Malone, J. Fostel, J. Zheng, J. Rees, L. Soldatova, L. Hunter, M. Brochhausen, M. Brush, M. Courtot, M. Dumontier, P. Ciccarese, P. Hayes, P. Rocca-Serra, R. Dipert, R. Rudnicki, S. Sahoo, S. Arabandi, W. Ceusters, W. Duncan, W. Hogan, Y. He. *IAO: Information Artifact Ontology*, https://github.com/information-artifact-ontology/IAO (accessed Oct **16**, 2021).
[42] R. R. Brinkman, M. Courtot, D. Derom, J. M. Fostel, Y. He, P. Lord, J. Malone, H. Parkinson, B. Peters, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, L. N. Soldatova, C. J. Stoeckert Jr., J. A. Turner, J. Zheng, The OBI Consortium. *J. Biomed. Semant.* **1**, S7 (2010).
[43] S. Tan, C. Mungall, N. Vasilevsky, N. Matentzoglu, D. Osumi-Sutherland, reality, W. Dahdul, K. Blumberg, J. Balhoff, Clare72, S. Robb, M. Haendel, P. L. Buttigieg, M. Harris, S. Köhler, R. Walls, diatomsRcool, A. Meier, R. Hoehndorf, euniceyi, G. Gkoutos, S. Bello. pato-ontology/pato: 2021-09-09 release (2021), https://doi.org/10.5281/zenodo.5499804.
[44] G. V. Gkoutos, P. N. Schofield, R. Hoehndorf. *Database* **2012**, bas033 (2012).

[45] J. Ison, M. Kalas, I. Jonassen, D. Bolser, M. Uludag, H. McWilliam, J. Malone, R. Lopez, S. Pettifer, P. Rice. *Bioinformatics* **2P**, 1325 (2013).
[46] H. Rijgersberg, M. van Assem, J. Top. *Semantic Web* **4**, 3 (2013).
[47] A.-L. Lamprecht, S. Naujokat, B. Steffen, T. Margaría (2010). Constraint-Guided Workflow Composition Based on the EDAM Ontology. arXiv. https://doi.org/10.48550/arXiv.1012.1640.
[48] N. Beard, F. Bacall, A. Nenadic, M. Thurston, C. A. Goble, S.-A. Sansone, T. K. Attwood. *Bioinformatics* **36**, 3290 (2020).
[49] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, M. Ashburner. *Nucleic Acids Res.* **36**, D344 (2008).
[50] J. Hastings, D. Magka, C. Batchelor, L. Duan, R. Stevens, M. Ennis, C. Steinbeck. *J. Cheminf.* **4**, 1 (2012).
[51] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes, C. Steinbeck. *Nucleic Acids Res.* **44**, D1214 (2015).
[52] C. T. Hoyt, C. Mungall, N. Vasilevsky, D. Domingo-Fernández, M. Healy, V. Colluru. *ChemRxiv* (2020), https://doi.org/10.26434/chemrxiv.12591221.v1, In preparation.
[53] C. Mungall, N. Vasilevsky. *chiro: CHEBI Integrated Role Ontology*, https://github.com/chiro (accessed Oct 16, 2021).
[54] Y. Djoumbou Feunang, R. Eisner, C. Knox, L. Chepelev, J. Hastings, G. Owen, E. Fahy, C. Steinbeck, S. Subramanian, E. Bolton, R. Greiner, D. S. Wishart. *J. Cheminf.* **8**, 61 (2016).
[55] J. Hastings, M. Glauer, A. Memariani, F. Neuhaus, T. Mossakowski. *J. Cheminf.* **13**, 23 (2021).
[56] J. Hastings, L. Chepelev, E. Willighagen, N. Adams, C. Steinbeck, M. Dumontier. *PLoS One* **6**, e25513 (2011).
[57] K. Eilbeck, S. E. Lewis, C. J. Mungall, M. Yandell, L. Stein, R. Durbin, M. Ashburner. *Genome Biol.* **6**, R44 (2005).
[58] Royal Society of Chemistry RSC Ontologies, https://github.com/rsc-ontologies (accessed Oct 16, 2021).
[59] K. Haug, K. Cochrane, V. C. Nainala, M. Williams, J. Chang, K. V. Jayaseelan, C. O'Donovan. *Nucleic Acids Res.* **48**, D440 (2020).
[60] P. Tremouilhac, C.-L. Lin, P.-C. Huang, Y.-C. Huang, A. Nguyen, N. Jung, F. Bach, R. Ulrich, B. Neumair, A. Streit, S. Bräse. *Angew. Chem. Int. Ed. Engl.* **59**, 22771 (2020).
[61] R. Kidd. *ICIC – 2008: The International Conference in Trends for Scientific Information Professionals; 19–22 Oct. 2008, Nice*, Infonortics, Nice (2008).
[62] C. Batchelor, T. Polajnar, R. Kidd. *2nd RSC-BMCS/RSC-CICAG Artificial Intelligence in Chemistry, 2 Sep–3 Sep 2019*, Royal Society of Chemistry, Cambridge, United Kingdom (2019).
[63] T. Millecam, A. J. Jarrett, N. Young, D. E. Vanderwall, D. Della Corte. *Drug Discov. Today* **26**, 19221928 (2021).
[64] H. Krieg. in *2021 Spring Allotrope Connect, Virtual Meeting*, Allotrope Foundation, Washington, DC, USA (2021), https://www.youtube.com/watch?v=BOOyHns-pqY (accessed Mar 15, 2022).
[65] E. Little, H. Krieg, T. Weber, G. Gross, J. Espejo, T. Reyes, K. Colsman. in *DRAFT/WIP: Allotrope Framework Term Curation Guide*, Allotrope Foundation, Washington, DC, USA (2018), https://gitlab.com/allotrope-open-source/allotrope-devops/-/wikis/uploads/Allotrope_Framework_Term_Curation_Style_Guide.docx (accessed Mar 15, 2022).
[66] W. Schafer, O. He, A. Dunn, Z. E. X. Dance. in *2021 Spring Allotrope Connect, Virtual Meeting*, Allotrope Foundation, Washington, DC, USA (2021), https://www.youtube.com/watch?v=HVv8TJc7p9c (accessed Mar 15, 2022).
[67] G. Mayer, A. R. Jones, P.-A. Binz, E. W. Deutsch, S. Orchard, L. Montecchi-Palazzi, J. A. Vizcaíno, H. Hermjakob, D. Oveillero, R. Julian, C. Stephan, H. E. Meyer, M. Eisenacher. *Biochim. Biophys. Acta* **1844**, 98 (2014).
[68] L. Montecchi-Palazzi, S. Kerrien, F. Reisinger, B. Aranda, A. R. Jones, L. Martens, H. Hermjakob. *Proteomics* **9**, 5112 (2009).
[69] G. Mayer, L. Montecchi-Palazzi, D. Ovelleiro, A. R. Jones, P.-A. Binz, E. W. Deutsch, M. Chambers, M. Kallhardt, F. Levander, J. Shofstahl, S. Orchard, J. A. Vizcaíno, H. Hermjakob, C. Stephan, H. E. Meyer, M. Eisenacher, HUPO-PSI Group. *Database* **2013**, bat009 (2013).
[70] D. Schober, D. Jacob, M. Wilson, J. A. Cruz, A. Marcu, J. R. Grant, A. Moing, C. Deborde, L. F. de Figueiredo, K. Haug, P. Rocca-Serra, J. Easton, T. M. D. Ebbels, J. Hao, C. Ludwig, U. L. Günther, A. Rosato, M. S. Klein, I. A. Lewis, C. Luchinat, A. R. Jones, A. Grauslys, M. Larralde, M. Yokochi, N. Kobayashi, A. Porzel, J. L. Griffin, M. R. Viant, D. S. Wishart, C. Steinbeck, R. M. Salek, S. Neumann. *Anal. Chem.* **90**, 649 (2018).