Automatic Ontology Extension for the ChEBI Ontology

Simon Flügel[1]

[1]Institute of Computer Science, Osnabrück University, Friedrich-Janssen Str. 1, 49076 Osnabrück, Germany

Abstract

In the life sciences, the amount of available knowledge has increased drastically over the last decades. Reference ontologies are an essential tool for organising and making this knowledge accessible. However, since they are designed and maintained manually, extending them is costly and keeping up with scientific progress is almost impossible. In this PhD project, we develop methods that automatically extend the coverage of reference ontologies while remaining faithful to developers' intentions, using the example of the Chemical Entities of Biological Interest (ChEBI) ontology.

In particular, we are interested in neural-symbolic integration methods that combine Machine Learning with axiomatic knowledge from the ontology. The research project focuses on three avenues: Firstly, we examine how to represent chemical structures in Machine Learning methods, in particular Graph Neural Networks. Secondly, axiomatising ontology classes in monadic second-order logic (MSOL) and first-order logic (FOL) and integrating them with OWL ontologies. And thirdly, we study the direct injection of ontology axioms into the training process of Machine Learning methods.

The overarching goal of this work is to provide ontology developers and domain experts with a suite of tools that lighten the load of manual ontology development and broaden the scope of reference ontologies without lowering quality standards.

Keywords

ChEBI, ontology extension, neural-symbolic integration, OWL

1. Introduction

Ontologies in the biomedical domain, such as those of the OBO Foundry [1], are maintained by manual curation. While this ensures high quality standards, it also limits the growth of ontologies. The ChEBI (Chemical Entities of Biological Interest) ontology [2] for example features 61,867 fully annotated compounds (as of version 242, released in June 2025). Comparing this with chemical databases such as PubChem [3], which has 121 million entries (as of July 2025), it becomes clear that ChEBI cannot not come close to a full coverage of its domain, at least not with manual curation alone.

Therefore, this research project focuses on developing and improving methods for automated ontology extension, i.e., the addition of content to ChEBI, or the application of ChEBI to new data.

A major challenge for ontology extension is staying consistent with the manually curated ontology. Each ontology term is the result of a consensus between experts and refers to background knowledge that is not always made explicit. This requires neural-symbolic techniques that can integrate knowledge from symbolic sources, e.g., the OWL axioms of the ontology, as well as sub-symbolic sources, i.e., the considerable amount of chemicals that are already annotated by ChEBI.

Given the diversity of the chemical domain, we believe there is no one-size-fits-all solution for ontology extension. Instead, our goal is to provide an ensemble of different approaches. In previous work, various Machine Learning methods have been applied to this task, including LSTMs [4] and Transformer models [5], the latter of which is performing best on the ontology extension task. While Transformer models are able to predict a large number of classes simultaneously (up to 1,332 [6]), they are by nature data-dependent. In the experiments, only classes with at least 50 or 100 molecules have been selected. Less populated classes, which make up the majority of ChEBI classes, have not yet been not covered. In a user study, it has been determined that a lack more specific classification is seen by

Proceedings of FOIS 2025 Satellite events co-located with the 15th International Conference on Formal Ontology in Information Systems (FOIS 2025), September 10-12, 2025, Catania, Italy

simon.fluegel@uos.de (S. Flügel)

0000-0003-3754-9016 (S. Flügel)

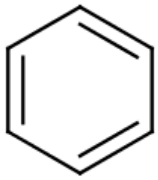

Figure 1: Benzene, a ring of 6 carbon atoms that can be represented in SMILES as *C=1C=CC=CC1*.

users as a major drawback [6]. Also, not all classes can be learned equally well. For instance, classes with complex ring structures pose a particular challenge [4].

This author has also participated in preliminary work on a logic-based approach to ontology extension using SMILES [7]. SMILES is a string representation language for molecules. Many ChEBI classes that represent groups of molecules have been annotated with their defining substructure. In order to use these substructures for ontology extension, SMILES strings have been translated into first-order logic (FOL) axioms which can then be used for classification [8].

In this project, we aim to answer the following research questions:

1. Can a graph representation of molecules and feature augmentation improve the performance of Machine Learning methods for ontology extension?
2. How can complex chemical definitions be formalised? And how can such formalisations be used for ontology extension?
3. Can ontology extension be improved by directly infusing the training process of Machine Learning methods with ontology axioms?

2. Molecule Representation

The first question that has to be addressed when using Machine Learning for the chemical domain is: How do we get molecular structures into a model? And which model provides the best representation for a molecule?

Fingerprints encode molecules as fixed-length vectors, taking the structure and physicochemical properties into account. Depending on the application, e.g., toxicity prediction or virtual screening, different types of fingerprints have been developed [9]. String representation such as SMILES or SELFIES [10] perform a traversal of the molecular graph, encoding each atom with a sequence of letters, with additional symbols for bonds, branches and ring structures.

In previous work, both fingerprints for classical Machine Learning methods and SMILES strings for Transformer models have been used [4, 5]. However, this representation is not optimal: Chemical structures come in many different shapes and sizes. Thus, fitting them into fixed-length or sequential representations makes it harder for a model to understand the original molecule. Take for instance benzene rings (cf. Figure 1), which can be described by the following SMILES string: *C=1C=CC=CC1*. In the structure, the first and the last *C* are direct neighbours, while in the representation, they are at opposite ends. Their connection can only be inferred from the ones that act as ring opening and closure symbols. While learning this dependency is not a problem for small ring sizes, larger rings with hundreds of atoms or complex ring structures pose a challenge for sequence-based models. Note that SMILES usually allows different text representations for the same molecule, For instance, *c1ccccc1* is also a valid description of a benzene ring. However, all face the same issue that they have to bring a circular structure into a linear form.

We hypothesise that a graph representation in which each atom is translated to a node and each bond becomes an edge would avoid such problems and facilitate the learning process. This representation

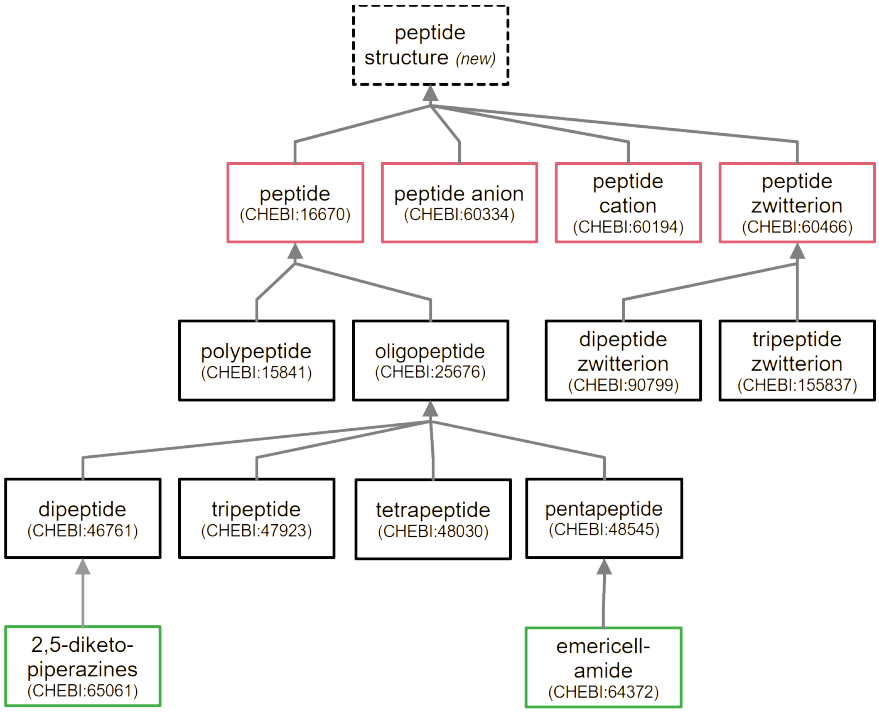

Figure 2: Peptide-related classes from ChEBI we have axiomatised in MSOL. *peptide structure* has been introduced as an artificial node into the hierarchy, capturing the common structural features of all classes. The colours correspond to different classification mechanisms: Red for charge-based, black for size-based and green for substructure-based classification.

also requires a new model architecture, namely Graph Neural Networks (GNNs) [11]. So far, preliminary experiments have been conducted which show that, for many ChEBI classes, GNNs perform better on the ontology extension task than Transformer models while using fewer resources. Our next steps will be to further optimise our GNN architecture using feature augmentation and a more sophisticated graph structure (e.g., with additional nodes for ring structures or functional groups). Also, we are going to investigate specialised pre-training methods for GNNs. This will result in an ensemble architecture where predictions from several models (including both Transformers and GNNs) are aggregated into a single classification.

3. Class Axiomatisation

In our work with ChEBI, we have identified several areas which are underspecified. For example, about 15 thousand compounds are classified as belonging to the *peptide* class. This class has a subhierarchy which allows for a more fine-grained classification, e.g., into *oligo-* and *polypeptides* (peptides with either "few" or "many" amino acids). These two subclasses, by their definition, form a partition of the peptide domain. This means that all peptide compounds could be classified into one of the two subclasses. However, there are 7,500 direct children of *peptide* which do not have an *is a* axiom to one of the subclasses. Such missing axioms make learning more difficult for Machine Learning models. For one, there is simply less data to train on, i.e., less positive samples for the classes *oligopeptide* and *polypeptide*. But there are also more negative samples, i.e., molecules that are chemically *oligo-* or polypeptides, but have a negative label in the Machine Learning dataset.

Therefore, in this project, peptides and 13 of its related classes or subclasses (cf. Figure 2) have been selected to study a rule-based approach to ontology extension [12]. In collaboration with domain experts, we have developed and refined natural language definitions for these classes. While natural-language definitions already exist in ChEBI for most classes, they do not cover some edge cases and presume chemical background knowledge. By formalising these definitions and developing a methodology for

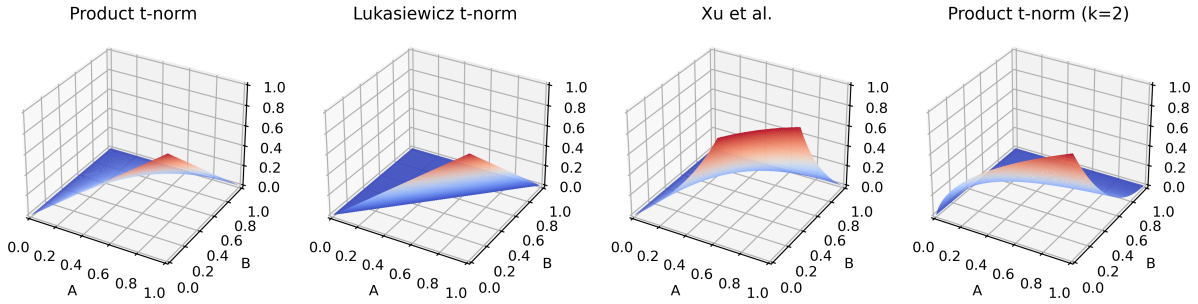

Figure 3: Loss for a subsumption relation $A \subseteq B$ given different prediction values for the classes A and B. Product and Łukasiewicz refer to the fuzzy t-norms, Xu et al. is a probabilistic semantic loss described in [14]. The right-most variant is a balanced fuzzy loss that gives the subclass more weight, counterbalancing the relative overrepresentation of superclasses in the dataset.

automatic classification, we were able to test our definitions and compare them against the current ChEBI classification. This has led to further refinement of our definition or, in some cases, to the identification of errors and inconsistencies in ChEBI.

The classification methodology that is necessary for this process is based on an MSOL axiomatisation. While [13] already identified MSOL as a suitable language for chemical class definitions (in their case, fullerenes), this project has determined that peptides require monadic second-order definitions as well. This is in contrast to OWL and FOL, which are more commonly used in ontology development, but in which we cannot express the concept of peptides.

This raises a new issue: How can one use second-order definitions in classification tasks? We address this issue with a methodology that translates the MSOL definitions into a FOL model checking problem in which some components are calculated algorithmically. Here, the central idea is that reasoning over the whole ontology is not necessary. Instead, each model checking problem has a single molecule as its domain. With this domain, which consists of the molecule's atoms and the algorithmically supplied components, model checking becomes feasible on the scale of ChEBI.

As a third step, the MSOL and FOL axiomatisations have been used to verify the trustworthiness of a purely algorithmic classifier. With the algorithmic method, we were able to classify all 121 million compounds of PubChem, which goes significantly beyond the scope of ChEBI.

In future work, we will generalise our methodology and extend it towards other areas of ChEBI.

4. Injecting Axioms into Training

One of the drawbacks of our current Machine Learning pipeline is that it is agnostic about the OWL axiomatisation of ChEBI. The only axioms that are used are the subsumption relations which connect label classes to compound classes. From this, the dataset is constructed, taking each compound as a sample with a list of positive or negative labels. Importantly, transitive parents are used as positive labels as well. Thus, given a peptide for instance, the model not only has to predict the peptide class, but also the amide class, the organic class and so on.

This ignores the subsumption relations between label classes (e.g., we can infer from the ontology that each peptide is organic and an amide). It also ignores disjointness and other relations between classes. Therefore, a model might predict that a molecule is a peptide and not an amide, or that it is both organic and inorganic. While it will still receive a loss for such predictions during training, this loss teaches the ontology axioms only by example, not as a general rule.

In [15], we introduced a *fuzzy loss* that combines a standard cross-entropy loss with additional loss terms based on a fuzzy logic interpretation of subsumption and disjointness axioms. This allows the model to learn relations between label classes directly instead of inferring them from samples. We evaluated the fuzzy loss for different fuzzy implications and parameter configurations (cf. Figure 3). Overall, we were able to improve the consistency of predictions significantly compared to a regular

cross-entropy loss.

In addition, we performed experiments with an additional unsupervised learning task. There, the fact is used that fuzzy loss works without labels: Even if the correct classification is unknown, we can tell if a classification is consistent or not. Data from PubChem was used to augment our labelled dataset with additional unlabelled samples. The goal was to improve the out-of-distribution generalisation abilities of the model by drawing data from a wider distribution than the original dataset.

The main drawbacks of the fuzzy loss are that, for one, in our experiments, adding fuzzy loss terms was detrimental to model performance. While a balancing technique reduces the performance gap between the baseline and fuzzy loss models, a successful application of the fuzzy loss would require further performance improvements on the classification task. Also, in [15], only subsumption and disjointness axioms have been taken into account, leaving out other axioms in ChEBI.

Expanding the fuzzy loss method to other axiom types poses a new challenge: How does the axiom correspond to the loss function? Unlike for subsumption and disjointness axioms, where this is relatively straight-forward (if A and B are disjoint, a model should not predict A and B at the same time), other axiom types are more complex. For instance, ChEBI uses the object property *has functional parent* for relations between classes where one can be derived from the other by functional modification. This does not give us clear rules we can use for a loss function. Given the knowledge that, lets say, *penicillin* has the functional parent *6-aminopenicillanic acid*, we can derive no statement about individual molecules that belong to either of the classes. One way to change that would be to train a separate model on the prediction of functional modifications, which has been relegated to future work.

For now, this project focuses on a promising axiom type that draws information from a different source: *has part* relations are used to relate classes to chemical parts which they contain. Since the chemical structures are provided by ChEBI, we can identify their parts and compare them to the axioms. Take for example the class *carboxylic acid*. It has an axiom *carboxylic acid* ***has part*** *some carboxy group*. If a model would predict a given molecule that has no carboxy group as a carboxylic acid, we would know that this prediction is wrong. We can identify the carboxy group based on the ChEBI class *carboxy group*, which, despite not being a molecule itself, is annotated with a SMILES string. This SMILES string describes the substructure we have to identify in the molecule.

While ChEBI already includes a subhierarchy for groups with more than 3,000 members, the parthood axioms linking them to molecule classes are relatively sparse. We aim to expand the coverage of ChEBI in this area, focussing specifically on the classes on which Machine Learning models are trained.

In a separate approach, we also plan to use parthood relations for feature augmentation. For individual samples, we can directly annotate molecules with the groups that are part of them. This does not require a class-level axiomatisation. Instead, the hypothesis is that the groups themselves are chemically relevant substructures. Since the groups in ChEBI have been selected by expert curators, giving this knowledge to Machine Learning models might induce them to learn concepts in a way similar to how experts usually see them, namely in terms of functional groups.

Finally, this project will investigate the use of Logical Neural Networks (LNNs, [16]), an architecture designed specifically to represent logic formulas. It is differentiable and captures logical contradiction directly in the loss function. Instead of a single prediction score, it outputs bounds which allow for a well-founded interpretation of predictions. For example, a huge gap between the lower and upper bounds represents uncertainty while a lower bounds that is higher than the upper bound indicates inconsistency. LNNs have the potential to provide a stronger integration between the ontology and sub-symbolic training and to yield more interpretable results.

Acknowledgments

This work has been funded by the Deutsche Forschungsgesellschaft (DFG, German Research Foundation) under grant number 522907718.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] R. Jackson, N. Matentzoglu, J. A. Overton, R. Vita, J. P. Balhoff, P. L. Buttigieg, S. Carbon, M. Courtot, A. D. Diehl, D. M. Dooley, W. D. Duncan, N. L. Harris, M. A. Haendel, S. E. Lewis, D. A. Natale, D. Osumi-Sutherland, A. Ruttenberg, L. M. Schriml, B. Smith, C. J. Stoeckert Jr., N. A. Vasilevsky, R. L. Walls, J. Zheng, C. J. M. Mungall, B. Peters, OBO foundry in 2021: Operationalizing open data principles to evaluate ontologies, Database 2021 (2021) baab069.

[2] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes, C. Steinbeck, ChEBI in 2016: Improved services and an expanding collection of metabolites, Nucleic Acids Research 44 (2016) D1214–D1219.

[3] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E. E. Bolton, PubChem 2025 update, Nucleic Acids Res. 53 (2025) D1516–D1525.

[4] J. Hastings, M. Glauer, A. Memariani, F. Neuhaus, T. Mossakowski, Learning chemistry: Exploring the suitability of machine learning for the task of structure-based chemical ontology classification, Journal of Cheminformatics 13 (2021) 1–20.

[5] M. Glauer, A. Memariani, F. Neuhaus, T. Mossakowski, J. Hastings, Interpretable ontology extension in chemistry, Semantic Web 15 (2024) 937–958.

[6] M. Glauer, F. Neuhaus, S. Flügel, M. Wosny, T. Mossakowski, A. Memariani, J. Schwerdt, J. Hastings, Chebifier: Atomating semantic classification in ChEBI to accelerate data-driven discovery, Digital Discovery 3 (2024) 896–907.

[7] D. Weininger, SMILES, a chemical language and information system, Journal of Chemical Information and Computer Sciences 28 (1988) 31–36.

[8] S. Flügel, M. Glauer, F. Neuhaus, J. Hastings, When one logic is not enough: Integrating first-order annotations in OWL ontologies, Semantic Web 16 (2025) SW–243440.

[9] J. Yang, Y. Cai, K. Zhao, H. Xie, X. Chen, Concepts and applications of chemical fingerprint for hit and lead screening, Drug discovery today 27 (2022) 103356.

[10] M. Krenn, F. Häse, A. Nigam, P. Friederich, A. Aspuru-Guzik, SELFIES: A robust representation of semantically constrained graphs with an example application in chemistry, arXiv preprint arXiv:1905.13741 1 (2019).

[11] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE Transactions on Neural Networks 20 (2008) 61–80.

[12] S. Flügel, M. Glauer, T. Mossakowski, F. Neuhaus, ChemLog: Making MSOL viable for ontological classification and learning, in: International Joint Conference on Learning and Reasoning, 2025, in submission.

[13] O. Kutz, J. Hastings, T. Mossakowski, Modelling highly symmetrical molecules: Linking ontologies and graphs, in: International Conference on Artificial Intelligence: Methodology, Systems, and Applications, Springer, 2012, pp. 103–111.

[14] J. Xu, Z. Zhang, T. Friedman, Y. Liang, G. Broeck, A semantic loss function for deep learning with symbolic knowledge, in: International Conference on Machine Learning, PMLR, 2018, pp. 5502–5511.

[15] S. Flügel, M. Glauer, T. Mossakowski, F. Neuhaus, A fuzzy loss for ontology classification, in: International Conference on Neural-Symbolic Learning and Reasoning, Springer, 2024, pp. 101–118.

[16] R. Riegel, A. Gray, F. Luus, N. Khan, N. Makondo, I. Y. Akhalwaya, H. Qian, R. Fagin, F. Barahona, U. Sharma, et al., Logical neural networks, arXiv preprint arXiv:2006.13155 (2020).