Q&A
13 October 2021
Automated Chemical Ontology Expansion using Deep Learning
Dr Janna Hastings (UCL)

Q1: When did you first start getting interested in using semantic web technologies and working with ontologies? And what sparked that interest?

I was hired many years ago by the EBI, the European Bioinformatics Institute, to work on ChEBI and before that I didn't really know about ontologies, so I was put in the deep end. I did study computer science and did study logic-based knowledge representation and in those days we didn't have OWL yet, so now we have OWL and all kind of different reasoners and all sorts of interesting things that we can do.

Q2: How robust are the machine learning based predictions for small changes in the class predictions?

If you have even a small change in the input structure, that might mean that it should be predicted to belong to a different class. Even small changes may have a big semantic impact on your classification. So, you would want it to be sensitive to the differences that matter and not sensitive to the differences that don't matter. That's the nuance of evaluation, so in order to do the evaluation, we used an unbiased sample of classes from ChEBI and took a test and training data split from the class selection in a totally randomized way, and we hope that therefore what we see in terms of performances, tracks really chemically interesting performance in an unbiased way.

Q3: Where can we find the code of the project?

I do not have the link right in front of me, but if you go to the paper you will find the GitHub link.

Q4: What kind of ontologies can this approach not be used for. For example, how dependent is it on a computer representation of the things that are to be categorized?

So, what this approach depends on is that there is some information annotated to the members that will be predictive of the class structure. So, it will work for types of ontology where you have a class structure that tracks something structural. In chemistry it's obvious what the structure is, it would work for proteins, and sequence classifications too. It would probably work for pathways which have some kind of structural information. It may work for text-type ontologies where the text where the entities are well described in some sort of text segment. But there are obviously lots of ontologies for which there's no annotation information that you can use for learning, and it won't work for those.

Q5: So, BioAssay's would be tricky then?

Probably, unless you have descriptions in text which have the information.

Q6: So I'm guessing this hasn't been tried with the BioAssay Ontology?

No

Q7: How is the performance of classification specifically for natural product compounds?

So, we didn't look at that as a different sub-group. ChEBI obviously has natural products as classes as well as less natural product classes. There is another work that I'm aware of, I've forgotten the authors name, but it's called NP classifier. And they are using deep learning to predict natural product classes specifically, so they don't look in the general case and they just look at list of a few hundred natural product classes and use a similar kind of deep learning approach there. And they found that their approach was working quite well, so I think you should start by looking at that. Search for "NP classifier", I think you'll find the paper that way.

Q8: Is the output of the learning task the association of chemical compounds to a single class, or multiple classes, or other structures of ChEBI? What else, other than class identification, are human curators curating?

The output of the learning task is associating of a compound to multiple classes. The human curators are checking all kinds of things and assembling all kinds of information. For example, they check the names of the chemicals for correctness, and they assemble things like trivial names, which you can't predict automatically. And, basically, this additional information, this particular learning task doesn't give you.

Q9: You mentioned the representation used for the classical machine learning training limited the prediction ability for certain descriptors. Which representation was used and was there a reason for this choice?

For the classical approaches we use the standard fingerprint from the RDKit which is a chemical informatics library in Python and the reason for that choice was, that was the most straightforward to get a straightforward representation that could be used for the learning task. And in fact, there are other fingerprints even within the RDKit, which would potentially contain more information, they are a little bit slower to compute, and this would potentially address the limitation for those classes like salts, which then lacked information to make predictions. So, finding the best fingerprint for the learning was a "we didn't try", we just use the simplest, quickest one.

Q10: Do you think that rule-based predictions (classifier) can be more robust?

There's no question that the rule-based prediction, with a rule usually if your antecedent is true, your consequent is true. So, you are sort of staying in territory where you are on firm ground and for sure you have some confidence, whereas with a learning approach it's a kind of statistical association that you're learning and therefore you can get really wrong predictions. But with the rule-based approach, for the most part, although you might not get wrong predictions, your predictions might be too general to be useful. So, getting the best prediction with a rule-based approach may involve hugely complex and perhaps impossible to maintain rules. Whereas, with a learning approach you can hope that at least a certain percentage of the time you get really the best prediction because that's what you've tried to learn. So, there's advantages and disadvantages.