chem-bla-ics

One Million IUPAC names

Egon Willighagen

Published March 8, 2025

Citation
Willighagen, E. (2025). One Million IUPAC names. In *chem-bla-ics*. chem-bla-ics. https://doi.org/10.59350/tjkf2-k1608

Keywords
Iupac, Cheminf, Oscar, Textmining, Europepmc

Abstract
Names of chemicals are part of the human user experience when browsing a chemical database. And literature too, of course. Chemical names are also not easy to use, and what a chemical name means is not always clear. This is why the IUPAC started a standardizing nomenclature in chemistry, the IUPAC names. Each IUPAC name uniquely defines the chemical structure it defines. For example, methane is the IUPAC name for the chemical CH4.

Names of chemicals are part of the human user experience when browsing a chemical database. And literature too, of course. Chemical names are also not easy to use, and what a chemical name means is not always clear. This is why the IUPAC started a standardizing nomenclature in chemistry, the *IUPAC names*. Each IUPAC name uniquely defines the chemical structure it defines. For example, *methane* is the IUPAC name for the chemical CH_4.

So, when propagating chemical structures from the Beilstein Bioschemas feed, I was looking for names, IUPAC or not, ideally the name used in the article. When I asked about this, the question came up if they could autogenerate IUPAC names, for which various new tools exist (I think I am missing one from an American team, but cannot find the reference), along with multiple established commerical tools. Because the IUPAC nomenclature is a long list of naming rules, priorities, etc, a rule-based algorithm is logical, but newer methods take a deep-learning approach.

Back to the chemical annotation of chemistry literature. This is of obvious interest: you want to know where we can read more about a certain chemical. We need the chemical structures in a database for that, linked to the articles. This is, of course, one of the original studies of *cheminformatics*. And when authors of the chemical literature do not provide this routinely (this post shows a few exceptions, but it is still all too rare). And then manual and automated curation is needed, e.g. done by Chemical Abstracts.

Third, Wikidata has about 1.4 million chemical compounds and many names. A property propoal for IUPAC names has been long pending, but once accepted in one form or another, will require IUPAC names too.

One million IUPAC names

Thus, the idea came up, can we create a set of 1 million unique IUPAC names found in literature? I asked on the ELIXIR Europe slack channel if Europe PMC had such a dataset (doi:10.1093/nar/gkad1085). I knew they had been adding chemical named-entity recognition (NER) results in their annotation API. I learned they used ChEBI. Melanie Vollmar and Summer Rosonovski or Europe PMC gave useful information and support. Magnus Palmblad also replied and provided Python code to use the Europe PMC API to fetch names it returns and see if those are IUPAC names. Well, that's easy. We have OPSIN for that (see doi:10.1021/ci100384d).

Unfortunately, the Europe PMC NER results are not ideal for IUPAC names. Just scanning some 5, 6 organic chemistry journals returned some 8 thousand IUPAC names in open access articles. But it quickly started to be too limited: each set of articles returned increasingly few new names. The reason is simple: the NER is too *greedy* and as a result, does not easily recognize longer IUPAC names. It is too happy with a substring of the IUPAC name. For example, when it encounters the IUPAC name *5-Bromo-1H-indole-3-carboxylic acid*, it settles for *indole-3-carboxylic acid*:

36098	salvage pathway for	pyrimidine	synthesis, our resu
36098	5-Bromo-1H-	indole-3-carboxylic acid	(2)
36098	oromethyl)phenyl)-1H-	indole	-3-carboxamide (3)
36098	romethyl)phenyl)-1H-	indole	-3-carboxamide (4a)
36098	romethyl)phenyl)-1H-	indole	-3-carboxamide (4b)
36098	romethyl)phenyl)-1H-	indole	-3-carboxamide (5a)
36098	romethyl)phenyl)-1H-	indole	-3-carboxamide (5b)
36098	romethyl)phenyl)-1H-	indole	-3-carboxamide (5c)
36098	romethyl)phenyl)-1H-	indole	-3-carboxamide (5d)
36098	idation of L-DHO to	orotate	, which is followed

Open-Source Chemistry Analysis Routines

During my PhD, in 2003, when I worked a few months with Prof. Peter Murray-Rust (University of Cambridge) and Prof. Janet Thornthon (EMBL-EBI), I learned about the research by Sam Adams (doi:10.1039/B411699M), Joe Townsend (doi:10.1039/B411033A), and Peter Corbett (doi:10.1007/11875741_11). One of the tools that used this research was (is) OSCAR, short for *Open-Source Chemistry Analysis Routines* (see this detailed write up by Peter MR). Later, in 2010 I visted Peter again, as postdoc, in Cambridge, and then worked on the OSCAR project too. And while OSCAR did a lot more, the integration of Corbett's NER research made OSCAR the obvious follow-up step in finding IUPAC names in literature.

And because OSCAR4 had been integrated into Bioclipse (doi:10.1186/1758-2946-3-41) and I had this ported to Bacting already (doi:10.21105/joss.02558), using this was trivial. The use of Europe PMC is different now, however, and we are no longer using the Annotations API, but just using it to find open access articles, and to get the full text in XML format. That allows a simple XPath search on <p> elements, pass the resulting string to OSCAR4, and the recognized names are checked with OPSIN. And with this approach, processing two of the five or six journals we earlier explored, we find another 40+ thousand IUPAC names. Quite a success, I am tempted to say.

A Blue Obelisk project

So, I started a new Blue Obelisk project, iupac-names, to collect 1M IUPAC names. For researchers to use, learn from, etc. Just IUPAC names. Not even the chemical structure, nor the link to the articles. The first is trivial to do with OPSIN, so the matching SMILES do not need to be stored. Links to literature is tricky because of the aforementioned issues, and we only want to know which (partial) IUPAC names occur in literature. If you really want to know in which articles that IUPAC name is found, you can simply do a search in Europe PMC.

And because we only store IUPAC names, this are very basic facts (this is an IUPAC name, as defined by OPSIN being able to generate a SMILES for this structure) and that that string occurs in some article) and we can share them as CCZero. We defined various milestones, and I am happy that the first two have been reached within two weeks:

- Milestone 10k (doi:10.5281/zenodo.14965762)
- Milestone 50k (doi:10.5281/zenodo.14978557)

This second milestone has 53848 unique names, but as literature goes, there are interesting variations, some likely because of typesetting leading to spaces added and missing. If we ignore

spaces and hyphens, we have 50534 names left (hence the milestone). But IUPAC names are also not fully unique, partly because of Unicode character variations and greek letter alternatives, and you may wonder how many different chemical structures this set reflects. While not perfect, the Standard InChI gives some lower limit, and we find 36528 InChIKeys in this second milestone.

Now, we need twenty times as much to reach the 1M IUPAC names, but given we have many, many more open access articles to process. The bottleneck seems to be mostly our workflow.

Can you contribute?

Yes, of course! This is an open science project. But please keep in mind the narrow focus of this project: only IUPAC names which can be found in (open access) literature. This project doed not accept autogenerated names (PubChem would have given use many millions already), nor IUPAC names from existing databases. Ideally, you are able to show the code you use to extract/find those names in literature.

Can I use these names?

First of all, this is what the CCZero license and open science nature of this project is about: reuse. We love to hear how you are using these names, tho, and we encourage you to write up how you are using them. You can use DataCite to cite the release you used, and citing this blog post by DOI is also possible.

Does it support my language too?

No, at this moment it only support IUPAC names in English. Dutch, French, Spanish, or Chinese IUPAC names are valid, but currently not supported. See also this post.

Will there be a publication?

Magnus and I intend so. We already submitted an abstract to the International Conference on Chemical Structures, which has a Collection in the Journal of Cheminformatics. If the abstract gets accepted, of course, we can submit there. Otherwise, we will look for another venue, likely diamond open access.

Where is your script?

Ah, fair point. We did not decide on the final license yet. I have used two scripts based on the template by Magnus. As soon as we have finalized the license, we will make those available.