chem-bla-ics

One Million IUPAC names #3: the 200 thousand milestone and 1 million IUPAC names

Egon Willighagen

Published June 9, 2025

Citation

Willighagen, E. (2025). One Million IUPAC names #3: the 200 thousand milestone and 1 million IUPAC names. In *chem-bla-ics*. chem-bla-ics. https://doi.org/10.59350/6f7he-kxt56

Keywords

Iupac, Textmining, Europepmc

I could not find the time earlier to report (reason), but three weeks ago we passed the fourth milestone release of the CCZero IUPAC names found in literature collection. This release contains 200026 IUPAC names, 168702 unique names, reflecting 116207 unique InChIKeys. Time for an update of the One Million IUPAC names project.

The current count actually is just above 230 thousand IUPAC names, but further growth may require new approaches, such as the four ideas I posted earlier. I have gone through all full-text Open Access articles provided by the Europe PMC API. Now, this list is not static, but I wanted to start using their bulk downloads anyway.

The current results

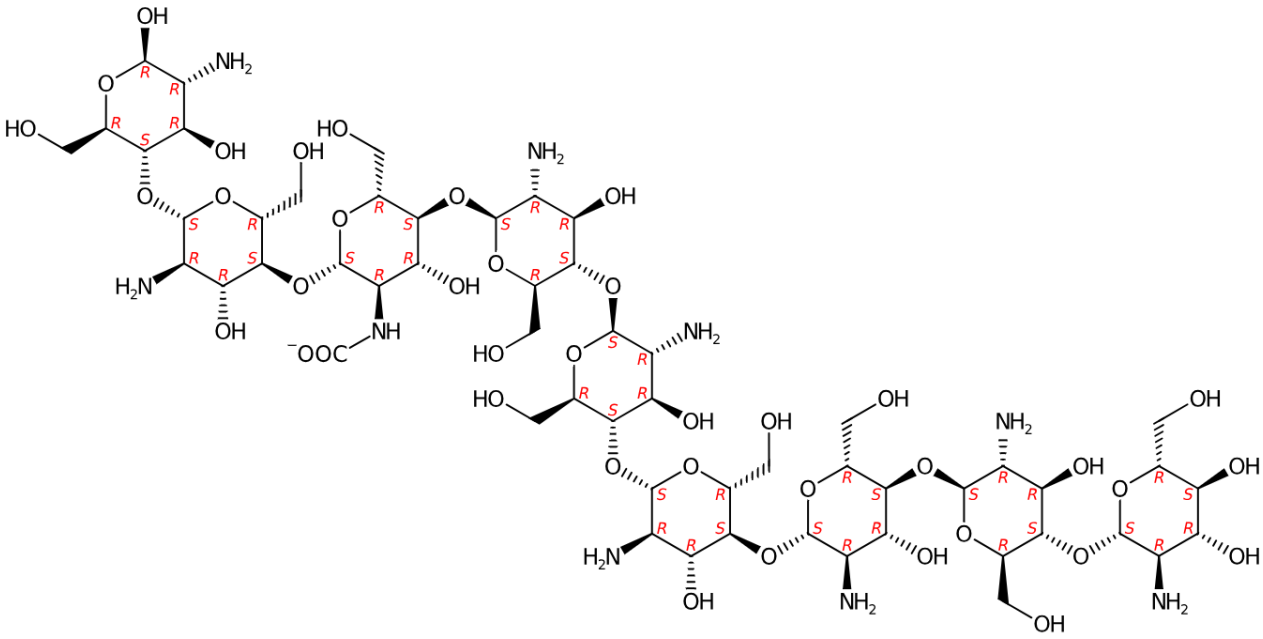
I have been looking at the names coming in. Some are short, others long. The complexity is fascinating and I will have to brush up my cheminformatics skills to make chemical space splots and visualize the structural diversity. I also note the current workflow does a good job at unicode characters, and we have plenty of names like `ε,ε-carotene-3,3'-dione`. There are also names that I do not expect to be really valid, like `hydroxymethyl methacrylate-` that end with a hyphen (41 in total), but their overall count is low. And OPSIN is happy with it, so the name fits the rules.

The ten longest names (so far) are these (with the lengths 322, 324, 332, 357, 371, 373, 376, 421, 429, and 626):

```
(5Z)-3-ethyl-5-[[4-[15-[7-[(Z)-(3-ethyl-4-oxo-2-sulfanylidene-1,3-thiazolidin-5-ylidene)m
(Z)-[[4-[[(Z)-N'-carbamoyl-N-[2-[2-[2-[[3-[(4S)-6,8-dichloro-2-methyl-3,4-dihydro-1H-isoc
2-((Z)-2-((6-(4-(6-((Z)-(1-(dicyanomethylene)-5,6-difluoro-3-oxo-1H-inden-2(3H)-ylidene)m
(2S,4S,5R,6R)-5-acetamido-2-[(2S,3R,4R,5S,6R)-5-[(2S,3R,4R,5R,6R)-3-acetamido-4,5-dihydro
(2R,3S,4R,5R,7S,9S,10S,11R,12S,13R)-7-[(benzylcarbamoyl)oxy]-2-(1-{[(2R,3R,4R,5R,6R)-5-hy
2-[4-[2-[[(2R)-1-[[(4R,7S,10S,13R,16S,19R)-10-(4-aminobutyl)-4-[[(2R,3R)-1,3 dihydroxybut
(2R,3S,4R,5R,7S,9S,10S,11R,12S,13R)-12-{[(2R,4R,5S,6S)-4,5-dihydroxy-4,6-dimethyltetrahyc
(2S,4S,5R,6R)-5-acetamido-2-[(2S,3R,4R,5S,6R)-5-[(2S,3R,4R,5R,6R)-3-acetamido-5-hydroxy-6
(2R,3S,4R,5R,7S,9S,10S,11R,12S,13R)-12-{[(2R,4R,5S,6S)-4,5-dihydroxy-4,6-dimethyltetrahyc
N-[(2S,3R,4R,5S,6R)-5-[(2S,3R,4R,5S,6R)-3-amino-5-[(2S,3R,4R,5S,6R)-3-amino-5-[(2S,3R,4R,
```

That last compound has the InChIKey `DKPKDPKJVDQUPD-XGBIXEJNSA-M` and cannot be found in Google nor in PubChem. It looks like this:

chem-bla-ics



There are some closely related compounds, though.

Chemicals only published about once

Some related data was blogged by Henry Rzepa last week, with this quote by Lee from CAS:

> *38.5% of the current substances have only 1 reference*

Apparently, based on CAS Registry data, about 1 in 3 chemical structures are only published about once. And two in three are published about at least twice. I agree with Henry here, with organic chemistry literature in mind, I would have expected that 38.5% to be higher.

Anyway, since this project is not tracking in which articles IUPAC names are found, I have nothing to study this.

1 million IUPAC names

So, the primary goal of this project is to reach one million IUPAC names. We are currently at around 23%. Not bad, considering we started in Februari. And we have plenty of untouched literature left.

But I also applied idea 1, the varying names. The idea is that this was I can explode the number of compounds. In that compounds above, just the number of variations by enumerating all `OH` replacements with `OMe` and `OEt` would help a lot.

Because I wanted to make sure I could answer positively at the ICCS if we made it to one million CCZero IUPAC names, I implemented a very simple enumeration script. Really dumb approach. But the results are interesting. I started with the 200026 names from the milestone. If I explode these names, I get 1,377,127 IUPAC names, well above the target. Even if I remove name variations due to unicode variations for hyphens, I still have 1,162,107 IUPAC names.

chem-bla-ics

Something interesting I cannot fully understand at this moment yet, however, is the following. When I calculate the number of unique InChIKeys for the milestone, I get 117,726 keys, and when I do this for the list of name variations, I get 203,979 keys. So, while the IUPAC name list is about five times as long, the list of InChIKeys is not even twice as long. Well, I guess that is why this is called research.