

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

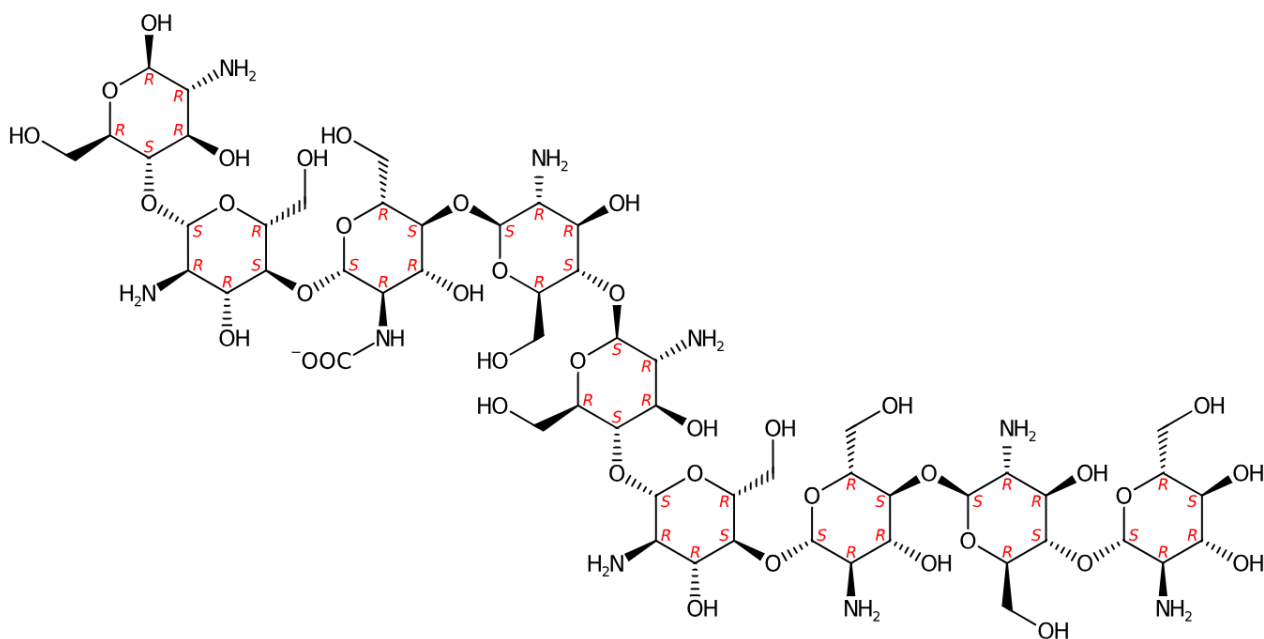
Abstract

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.

Copyright

Copyright © Egon Willighagen 2025. Distributed under the terms of the [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

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.