chem-bla-ics

# Creating Preferred IUPAC names

**Egon Willighagen**

Published November 1, 2026

**Citation**
Willighagen, E. (n.d.). In *chem-bla-ics*. chem-bla-ics. https://doi.org/10.59350/v7f5q-e2t67

**Keywords**
Iupac, Cheminf, Opensource, Openscience

**Copyright**
Copyright © Egon Willighagen 2026. Distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Language models model languages. The simpler the language, the simpler the model. Large language models (LLMs) do more: they do not merely capture a language, but also model what you can say with them. And you can map one model to another, and use that to translate things between langauges. Have a numerical representation, and you can have your LLM run on a processing unit (CPU, or more likely, GPU). More and more models are available (mostly proprietary, commercial) to use this. Chat bots, translation tools, grammar checkers, generative models. And because of the numerical representation, it is also easy enough to combine language with numerical data.

And I guess converting a SMILES to something else should not surprise me. Using SMILES as molecular descriptor has been around for some time. Using a generative model to generate an IUPAC name is another example. STOUT had been doing that for some time (doi:10.1186/s13321-021-00512-4). But based on deep learning, it is based on probabilities. Or recovering SMILES strings from Morgan fingerprints (doi:10.59350/srwee-de037).

The alternative is to generate IUPAC names using the IUPAC rules. Computational tools that do this have been around for a long time. The IUPAC naming rules are defined in the blue books, see this page. At some point in the nineties, based on the work of the *Woordenboek Organische Chemie*, I proposed to make a HTML version of the Dutch translation. The never happened.

Now, these LLMs have become quite powerful that they can do a ton of things. They can without problem use external tools. That makes a lot of sense: why would you predict something when you can look it up or easily computer. That helps with exact validation. For example, Claude is said to validate JSON to make sure it never outputs JSON that is not valid (which it does without that validator).

So, why not translation the IUPAC rules into a program and use that program to deterministically calculate IUPAC names? Or even Preferred IUPAC names (PINs)? Wikidata wants those. The lack of a tool to determine if a name is a PIN or to just calculate it blocked the addition of a Wikidata property earlier this year.

Noel O'Boyle (now EMBL-EBI) found and compared three open source tools, each with a matching preprint: NISPO (doi:10.48550/arXiv.2607.26113), OpenClatura (doi:10.26434/chemrxiv.15006114/v2), and Orthonym (doi:10.26434/chemrxiv.15009769/v1). They yield 85-95% valid IUPAC names, as reported in Noel's blog post. Of these, Orthonym by default returns PINs and then the yield is a good bit lower.

A fourth tool is mentioned in the blog: smiles-to-iupac, but no compared. The author of this tool, Inwan Yoo, has been actively contributing to the OPSIN tool.

There is much to like about this, even the "competition". These projects will be compared. Where the differ, it will highlight where all tools have it wrong, some of the tools, and even none of the tools, and the IUPAC rules turn out to be "wrong" (more likely, ambiguous). The same for OPSIN (doi:10.1021/ci100384d). I think this is a win for the open science chemistry community and looking forward to what will happen in the next few months.

## chem-bla-ics

And, I will need to find some time to see how many Wikidata SMILES I can convert into PINs.

### References


Lowe, D. M., Corbett, P. T., Murray-Rust, P., & Glen, R. C. (2011). Chemical Name to Structure: OPSIN, an Open Source Solution. *Journal of Chemical Information and Modeling*, *51*(3), 739–753. https://doi.org/10.1021/ci100384d **[cito:citesAsAuthority]** CrossRef Scholia

Mirza, A., Jablonka, K. M., & Fedorov, R. (2026). *Openclatura–an Open-Source Nomenclature Framework for Rule-Based Molecule Naming*. American Chemical Society (ACS). https://doi.org/10.26434/chemrxiv.15006114/v2 **[cito:citesAsPotentialSolution]** CrossRef Scholia

Rajan, K., Zielesny, A., & Steinbeck, C. (2021). STOUT: SMILES to IUPAC names using neural machine translation. *Journal of Cheminformatics*, *13*(1). https://doi.org/10.1186/s13321-021-00512-4 **[cito:citesForInformation]** CrossRef Scholia

Rajan, K., Zielesny, A., & Steinbeck, C. (2026). *Orthonym: verified IUPAC names for chemical structures in the wild*. American Chemical Society (ACS). https://doi.org/10.26434/chemrxiv.15009769/v1 **[cito:citesAsPotentialSolution]** CrossRef Scholia

Runcie, N. T., Imrie, F., & Deane, C. M. (2026). *NISPO: Open-source IUPAC name generation tool* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2607.26113 **[cito:citesAsPotentialSolution]** CrossRef Scholia

Swain, J. (2026). Morg2SMILES: Generative Recovery of SMILES Strings from Morgan Fingerprints. In *Front Matter*. Front Matter. https://doi.org/10.59350/srwee-de037 CrossRef Scholia