

# Retrieving PFAS Lists on PubChem with PUG REST

Emma L. Schymanski<sup>1</sup>, Paul A. Thiessen<sup>2</sup>, Jeff Zhang<sup>2</sup> and Evan E. Bolton<sup>2</sup>

13/03/2022

<sup>1</sup> Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 avenue du Swing, 4367, Belvaux, Luxembourg. ELS: ORCID 0000-0001-6868-8145

<sup>2</sup> National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, 20894, USA. PAT: ORCID 0000-0002-1992-2086, JZ: ORCID 0000-0002-6192-4632, EEB: ORCID 0000-0002-5959-6190

## Background

This document is based off code developed at BioHackEU2020 arising from PubChemLite efforts [1]. The original efforts have a dedicated keyword folder and more extensive documentation within the public ECI GitLab pubchem repository.

PubChem [2] Classification Trees are here and accessed by their respective `hid` number (see Figure 1).

The screenshot shows the PubChem Classification Browser interface. At the top, there's a navigation bar with the NCBI logo and a search bar. Below the navigation bar, the title "PubChem Classification Browser" is displayed, along with a "Help" link. A brief description of the browser's purpose is provided. The main section is titled "Select classification" and "Search selected classification by". It features a dropdown menu for "NORMAN Suspect List Exchange" and a search input field with a "Search" button. Below this, a "Classification description" section explains the NORMAN Suspect List Exchange (NORMAN-SLE). Further down, there are options for "Data type counts to display" (None, Compound, Substance) and "Display zero count nodes?" (Yes, No). The bottom section, "Browse NORMAN Suspect List Exchange Tree", shows a hierarchical tree structure. The root node is "NORMAN Suspect List Exchange Classification" with a count of 112,236. It branches into several sub-nodes, each with a count: S13 | EUCOSMETICS | Combined Inventory of Ingredients Employed in Cosmetic Products (2000) and Revised Inventory (2006) (3,850), S25 | OECDPFAS | List of PFAS from the OECD (3,677), S36 | UBAPMT | Potential Persistent, Mobile and Toxic (PMT) substances (254), S50 | CCSCOMPEND | The Unified Collision Cross Section (CCS) Compendium (630), and S60 | SWISSPEST19 | Swiss Pesticides and Metabolites from Kiefer et al 2019 (1,343).

Figure 1: *Figure 1: The NORMAN Suspect List Exchange PubChem Classification Browser.*

A set of base functions were developed (ELS, PAT, EEB) based on largely undocumented features of the classification trees. These are included in the `hid_tree_JSON.R` and `hid_tree_functions.R` scripts. The latter is the home for all eventual functions (including those developed during BioHackathon and beyond), with a contents listing. The functions are documented using comments above the respective function, but not yet in `roxygen`.

To start, set up all packages, download latest scripts and source them:

```
library(rjson)
library(httr)
```

```
## Warning: package 'httr' was built under R version 3.6.3
```

```
hid_fn_file_url <- "https://gitlab.lcsb.uni.lu/eci/pubchem/-/raw/master/annotations/keywords/hids/hid_t
download.file(hid_fn_file_url,"hid_tree_functions.R")
source("hid_tree_functions.R")
```

### Retrieving one tree file

First, try to retrieve the CompTox tree, hid=105:

```
tree_csv <- getPcHidTree(105,3)
tree_csv
```

```
## [1] "classification_tree_hid105_depth3_export.csv"
```

Then, load & take a look:

```
CompTox_tree <- read.csv(tree_csv, stringsAsFactors = F)
```

For this application, we are most interested in the PFAS entries.

```
i_pfas_lists <- grep("PFAS",CompTox_tree$nodeNames)
CompTox_pfas_lists <- CompTox_tree[i_pfas_lists,]
```

Once we have this output, including node HNIDs, we can build the URL needed to retrieve the CID listings per node entry via PUG REST.

```
# https://pubchem.ncbi.nlm.nih.gov/rest/pug/classification/hnid/<integer>/<id type>/<format>
hnid_base_url <- "https://pubchem.ncbi.nlm.nih.gov/rest/pug/classification/hnid/"
hnid_end_url <- "/cids/TXT"
CompTox_pfas_lists$REST_URL <- ""

for (i in 1:length(CompTox_pfas_lists$nodeHNID)) {
  CompTox_pfas_lists$REST_URL[i] <- paste0(hnid_base_url,CompTox_pfas_lists$nodeHNID[i],hnid_end_url)
}
```

Note that since the CompTox tree has a nested format, there is a lot of duplication. Thus, the following trims two columns and uniquifies over the remaining entries:

```
CompTox_pfas_lists <- CompTox_pfas_lists[,c(-8,-9)]
CompTox_pfas_lists <- unique(CompTox_pfas_lists)
```

Now, write output and use this to update the *PFAS List of Lists*:

```
write.csv(CompTox_pfas_lists,file="CompTox_PFAS_lists_in_PubChem.csv",row.names = F)
```

To be continued ...

### References

1. Schymanski EL, Kondić T, Neumann S, et al (2021) Empowering large chemical knowledge bases for exposomics: PubChemLite meets MetFrag. Journal of Cheminformatics 13:19. <https://doi.org/10.1186/s13321-021-00489-0>

2. Kim S, Chen J, Cheng T, et al (2021) PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Research* 49:D1388–D1395. <https://doi.org/10.1093/nar/gkaa971>