# Motivation

I thought of organizing the file into folders based on family\_name. The reason is because the analytical chemist, in this case, could work with plant-based samples with the need of differentiating and recognizing the different compounds. Thus we could divide our molecular compounds according to the family of plants they come from. This choice was supported by the fact that I checked through the method sort\_values() how many folders would need to be created in each different kind of partitions. The category “family\_name” would imply a partition in 68 folders, which I considered not too many nor too little, compared to the partition through “org\_name” for example.

# Data cleaning

I though of creating a Dataframe, called “df”, from the name of the files inside the folder mol files. The first column (“df[‘Molecule’]”) of this Database would correspond to the name of .mol file itself and the second column (“df[‘Cid’]”) would correspond to the “pubchem\_cid”, which was contained in the final part of the name of the .mol files delimited by “\_”.

In two .mol files the “pubchem\_cid” was delimited by “-“ instead of “\_”, so I renamed those two files accordingly. In particular, I changed the names:

((1S,3Ar,5S,6R,7S,7Ar)-1-((R)-1-Acetoxyethyl)-6-(2-Methylbutanoyloxy)-4-Methylene-771717007

into

((1S,3Ar,5S,6R,7S,7Ar)-1-((R)-1-Acetoxyethyl)-6-(2-Methylbutanoyloxy)-4-Methylene\_771717007

and

((1R,5S,6R,8S,8Ar)-8-Acetoxy-5-((2S,3As,5R,6Ar)-5-Hydroxyhexahydrofuro[2,3-B]Furan-2-Yl)-44566824

into

((1R,5S,6R,8S,8Ar)-8-Acetoxy-5-((2S,3As,5R,6Ar)-5-Hydroxyhexahydrofuro[2,3-B]Furan-2-Yl)\_44566824

# Data Analysis

From the NPASS csv file 'data\_catalogue\_NPASS\_plant\_compounds.csv' I extract a Dataframe (‘npass’) using “|” as the separator for each column, like in the csv file. Then, I create another Data frame “da” out of the ‘npass’ Dataframe with all the rows which have a ‘pubchem\_cid' corresponding to the ‘Cid’ number of the Dataframe “df”, and removing all the duplicates. After this first skimming, we are left with 4005 rows. Many of these rows should be removed, as the original .mol files were only 309. This means that there are files which rows that have the same “pubchem\_cid” value but different other values like “np\_id”, “org\_id” or “family\_name” for example.

At this point, I checked inside the .mol files looking for information which helped to me to distinguish between the elements with the same “pubchem\_cid”, but I could not, as there is mainly information about mass in there. Unfortunately, there is no similar information listed in the NPASS file.

Then, I took the decision of dropping the duplicates (keeping the first) with the same “np\_id”, “pref\_name” and I add another column “family\_name” to the Dataframe “df” listing the family name for each molecule, as explained above in Motivation.

Then, using the tools of the os and shutil libraries I create the folders accordingly and move the files .mol in their respective folders.