Subject: Sorting, counting and deleting different elements (e.g., Iodine) in a dataset
Posted by Jo W on Sun, 11 Sep 2022 22:54:09 GMT
View Forum Message <> Reply to Message

How do you find out the distribution and numbers of different elements that might occur in a dataset of drug-like molecules and delete those compounds that contain these specific elements?

For example if you download 5000 HIV active organic compounds from say PubChem containing a diverse set of different structures, there will be some compounds that for example contain selenium atoms or iodine.
These type of elements are not common in many datasets for biological screening and can distort and/or cause poor model predictions to occur.
So, how can you collate these compounds in Datawarrior and quickly analyse their frequency and also selectively remove them?

I know you can set up a filter for example "molecular formula" or "smiles" and then type in "Se" and then the filter "hides" all the selenium containing compounds (if you reverse the filter) and you can also tell how many compounds in the dataset were "hidden" and therefore get a figure of the selenium-containing compounds in the dataset.

However it's very laborious to do this for all other elements (accepting that you want for example C,H,O,N elements to remain), and also this peace-meal approach does not let you visualise the number of compounds in the dataset that contain for example, selenium, iodine, chlorine, phosphorous, etc.
For example, it would be good to see the following as a table in DW:
C,H,N,F - 90 compounds
C,H,O,P - 200 compounds
C,H,O,Se - 10 compounds etc
and maybe to visualise them as a histogram. Then for example removing the selenium-containing compounds from the dataset, to see what effects on the model they have.
So how can this be achieved in DW?