Cheminformatics – programming assignment #1

Intro to RDKit and similarity search

In this assignment, we are going to explore the possibilities of one popular and nowadays widely used Python library for cheminformatics, the so-called RDKit. We are about to apply a few functions on one interesting dataset, which consists of „Levodopa“ like compounds encoded by the SMILES strings (recall your understanding from the first lecture).

Subtasks to be graded (1 pt each):

1. load dataset in a Jupyter notebook and extract relevant column to be processed

2. compute molecular fingerprints for each compound

3. compute Tanimoto similarity measure for each pair of compound x reference compound

4. provide sorting, report top 3 hits and discuss your results (with reference to the first lecture and limitations)

5. publish your notebook to a personal GitHub repo and share it with me

REFERENCE COMPOUND: LEVODOPA

C1=CC(=C(C=C1CC(C(=O)O)N)O)O

https://pubchem.ncbi.nlm.nih.gov/compound/6047

What is Levodopa and some reading about its forms

https://www.parkinson.org/Understanding-Parkinsons/Treatment/Prescription-Medications/Levodopa

If you prefer a study resource in czech, here we go https://www.wikiskripta.eu/w/Antiparkinsonika

Prerequisites

• Python3 installed on your machine

• RDKit library

• Jupyter notebook/lab

Deadline: in two weeks

BONUS: implement N similarity measures (such as Tanimoto, Dice, Euclidean) x M (fingerprints) and compare them in a graphical form