**Raport wstępny – Temat 9: Klastrowanie publicznych baz związków**

1. **Link do repozytorium kodu:** <https://github.com/JKosciukiewicz/MLDD_Project>
2. **Dane, które można wykorzystać do realizacji projektu:**

**-** Bazy MCule;

**-** Bazy ChEMBL;

**-** Bazy ZINC;

- Bazy PDB;

- Wstępnie przygotowanych zbiorów danych aktywnościowych.

Z dużym prawdopodobieństwem nasza praca będzie oparta na zbiorach bazy ChEMBL, gdzie znajdują się dane aktywnościowe związków.

1. **Narzędzia potrzebne do realizacji projektu:**

- Conda – umożliwi zarządzanie złożonymi zależnościami i różnorodnymi środowiskami prac, bibliotekami oraz uruchomienie i aktualizację pakietów języka Python;

- GitHub - przechowywanie kodu, śledzenie zmian w kodzie, zarządzanie różnymi wersjami projektu.

*Pakiety, biblioteki:*

- Scikit-learn - testowania modeli k-means na różnych zestawach danych, wykorzystanie algorytmu klastrowania k-means, aby podzielić dane na grupy; Biblioteka scikit-learn zasadniczo zorientowana na obliczenia na CPU.

- RDKit **-** narzędzie do generowania fingerprintów Morgana.

- PyTorch – biblioteka do Pythona, która pozwala budować modele sieci neuronowych do wykorzystania w klasteryzacji.

*Narzędzia:*

Analiza może opierać się o różne fingerprinty strukturalne, np. z wykorzystaniem:

- PCA do wizualizacji przestrzeni chemicznej;

- t-SNE wizualizacja danych w dwuwymiarowej przestrzeni.

1. **Problemy związane z realizacją projektu:**

Publiczne bazy danych są bogate w liczne związki chemiczne, jednak ich jakość często bywa niewystarczająca. Problemem jest publikowanie głównie pozytywnych wyników, co skutkuje pominięciem wielu związków nieudanych w testach laboratoryjnych. Ponadto, różnice w sprzęcie używanym w różnych ośrodkach naukowych przyczyniają się do niespójności danych.