**Objective:**

Create a basic **machine learning model** using bioactivity data from the **ChEMBL database** to predict the biological activity of chemical compounds. This prototype demonstrates data retrieval, preprocessing, and model building.

**Steps:**

1. **Install Required Libraries:** Install the necessary Python libraries to interact with the ChEMBL database and build the machine learning model.

A close up of a text

Description automatically generated

1. **Retrieve Bioactivity Data from ChEMBL:** Use the ChEMBL web service client to query and retrieve data for a specific target (e.g., a protein involved in a disease).

A screenshot of a computer

Description automatically generated

1. **Data Preprocessing:** Convert the retrieved data into a pandas DataFrame and clean it for machine learning tasks. This includes handling missing values, filtering by standard units, and converting data types.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. **Feature Selection:** For simplicity, select features like molecular weight, logP, and other molecular descriptors from the compounds.

A screenshot of a computer

Description automatically generated

1. **Model Building:** Split the data into training and testing sets, and build a basic machine learning model (e.g., a decision tree) to predict bioactivity.
2. **Evaluation:** The model's performance is evaluated based on the mean squared error (MSE) metric. A lower MSE indicates better predictive accuracy.

**Outcome:**

This basic prototype demonstrates the ability to retrieve and preprocess biological activity data, select relevant features, and build a simple machine learning model. It proves the concept of using data science for computational drug discovery, showing how bioactivity data can be leveraged to predict the effectiveness of new compounds.