

## Assignment 3: Exploratory Data Analysis & Descriptor/Fingerprint Calculation

Name: Michelle Mupanduki

Target Protein: EGFR (Epidermal Growth Factor Receptor)

Dataset: 66 molecules from ChEMBL database

### Key EDA Findings

#### Dataset Overview

Original dataset: 81 molecules

After cleaning (removing NA values and intermediate class): 66 molecules

Bioactivity classification: 20 active ( $\text{pIC}_{50} \geq 6$ ), 46 inactive ( $\text{pIC}_{50} < 6$ )

#### pIC<sub>50</sub> Statistics

Statistic	Value
Mean	5.12
Std	1.49
Min	3.00
Median	4.90
Max	8.80

#### Key Observations from Visualizations

The pIC<sub>50</sub> distribution histogram showed a right-skewed distribution, with most compounds clustering between pIC<sub>50</sub> 3–5, indicating a majority of weakly active or inactive compounds.

The bioactivity class bar plot confirmed the dataset is imbalanced, with inactive compounds (~46) outnumbering active compounds (~20).

The pIC<sub>50</sub> boxplot showed a clear separation between active (median ~7) and inactive (median ~4.5) compounds, confirming pIC<sub>50</sub> is a strong discriminator of bioactivity.

The MW vs LogP scatter plot showed active and inactive compounds overlap in chemical space, with most compounds falling within Lipinski's drug-likeness range (MW 200–600, LogP 2–5).

### Lipinski Descriptor Statistics (Mann-Whitney U Test)

Descriptor	Statistics	p-value	Interpretation
pIC <sub>50</sub>	920.0	$1.308 \times 10^{-10}$	Significant (reject H <sub>0</sub> )
NumHAcceptors	743.5	0.000056	Significant (reject H <sub>0</sub> )
LogP	588.0	0.075	Non-significant (fail to reject H <sub>0</sub> )
MW	529.0	0.339	Non-significant (fail to reject H <sub>0</sub> )
NumHDonors	425.5	0.618	Non-significant (fail to reject H <sub>0</sub> )

Statistically significant descriptors: pIC50 and NumHAcceptors showed significant differences between active and inactive compounds ( $p < 0.05$ ).

Non-significant descriptors: MW, LogP, and NumHDonors showed no statistically significant difference between active and inactive compounds.

## 2D Descriptor Statistics

Tool: PaDELpy

Number of 2D descriptors generated: 881

## **Fingerprint Calculation**

Tool Used

PaDELpy (Python wrapper for PaDEL-Descriptor software)

Fingerprints Calculated

Fingerprint Type	Bits/Features	Output File
PubChem Fingerprints	881 bits	pubchem_fingerprints.csv
Substructure Fingerprints	307 bits	Substructure_fingerprints.csv

PaDELpy was selected for the following reasons:

It is an open-source, Python-compatible tool ideal for Google Colab workflows.

It provides a comprehensive set of molecular descriptors (1,444 2D, 431 3D) and multiple fingerprint types (PubChem, MACCS, Substructure, etc.) in a single package.

PubChem fingerprints were specifically chosen because they are widely used in QSAR modeling and capture a broad range of structural and chemical features relevant to drug discovery.

Substructure fingerprints were additionally calculated as they encode specific functional group patterns, which are particularly relevant for EGFR inhibitor analysis.

PaDELpy integrates seamlessly with pandas DataFrames, making downstream ML model preparation straightforward.

## **Files Generated**

df\_lipinski.csv — Lipinski descriptors

pubchem\_fingerprints.csv — PubChem fingerprints

Substructure\_fingerprints.csv — Substructure fingerprints

QSAR\_dataset.csv — Combined ML-ready dataset

mannwhitney\_summary.csv — Mann-Whitney U test results

Plots: histogram, barplots, boxplots, scatter plot