Assignment Summary: Exploratory Data Analysis and Descriptor/Fingerprint Calculation

Name: Philippa Jennifer Ayiku

Exploratory Data Analysis (EDA) - Key Findings

In this study, we performed exploratory data analysis and descriptor calculation for a set of bioactive compounds to prepare a dataset for QSAR modelling. After cleaning the initial dataset of 29 compounds by removing missing values and intermediates, 10 compounds remained, with duplicate SMILES aggregated using the median IC50 values. IC50 values were then transformed to pIC50 (-log10 IC50 in M) to normalize the data, and compounds were classified as active ($pIC50 \geq 6$) or inactive ($pIC50 < 6$). Visualizations, including histograms, bar plots, and boxplots, highlighted the distribution of bioactivity and chemical properties, while scatterplots of molecular weight versus LogP provided insight into the chemical space of the dataset with size proportional to pIC50. Statistical analyses using the Mann-Whitney U test identified which descriptors (pIC50, MW, LogP, NumHDonors, NumHAcceptors) significantly differed between active and inactive compounds. Significant p-values (<0.05) indicate descriptors differ between active and inactive compounds.

Lipinski's Rule of Five and 2D Descriptors

To characterize the drug-likeness of the molecules, four key 2D descriptors (molecular weight, LogP, number of hydrogen bond donors, and number of hydrogen bond acceptors) based on Lipinski's Rule of Five were calculated using RDKit. The majority of compounds conformed to Lipinski's criteria, indicating favourable pharmacokinetic properties. Boxplots of these descriptors revealed trends that distinguish active from inactive compounds, supporting their relevance for QSAR modelling.

Fingerprint Calculation

Molecular fingerprints were calculated using PaDEL-Descriptor, a widely used tool for generating standardized chemical fingerprints. PubChem fingerprints (881-bit binary vectors) and substructure fingerprints (307-bit vectors) were generated to capture the structural features of each compound. This method was chosen for its reproducibility, support for multiple

fingerprint types, and seamless integration with activity data. These descriptors and fingerprints were combined with meta-information such as compound ID, bioactivity class, and pIC50 to create a comprehensive dataset ready for machine learning applications.

Overall, this analysis provided a clear view of the chemical space, key molecular descriptors, and structural features of the compounds, laying a robust foundation for QSAR modelling and predictive analysis of bioactivity.