Target Protein: Breast cancer type 1 susceptibility protein (BRCA1)

Dataset Overview

- Number of molecules: 16
- Bioactivity classes: 14 inactive, 2 active
- pIC50 statistics:
    - Mean: 4.50
    - Min: 3.60
    - Max: 6.51
    - Std. Dev.: 0.85

Key EDA Findings with Visualizations The dataset is highly imbalanced (14 inactive vs. 2 active), with most compounds showing weak or no activity against BRCA1 (pIC50 mostly 3.6–4.8; only two compounds exceed pIC50 6.0).

Main visualizations and insights:

- **Bar plot of bioactivity class counts** Inactive: 14 | Active: 2 → Strong class imbalance limits robust statistical comparisons.
- **Scatter plot: MW vs LogP (colored by class)** Active compounds cluster at higher LogP values (around 0 to +2), while inactive ones are mostly hydrophilic (LogP < 0). MW shows little separation.
- **Boxplots of Lipinski descriptors by class**
    - **LogP**: clearest trend — active compounds have significantly higher (less negative) LogP (median ~0 vs. ~−2 for inactive). Suggests increased lipophilicity favors activity.
    - **MW**: no meaningful difference (medians ~560–565 Da, heavy overlap).
    - **NumHDonors**: slight decrease in active (median ~6.5 vs. ~7), but overlap.
    - **NumHAcceptors**: lower in active (median ~6.5 vs. ~8), consistent with simpler/polar-group-reduced structures.
- **Histogram/Density of pIC50** (from notebook) Distribution is right-skewed with most values clustered below 5, and a small tail of more potent compounds.

Lipinski and 2D Descriptor Statistics Calculated using RDKit (from df_lipinski.csv): MW, LogP, NumHDonors, NumHAcceptors.

- Mann-Whitney U test (active vs. inactive):
    - Only pIC50 itself significant ($p = 0.0308 < 0.05$).
    - All descriptors non-significant ($p > 0.05$): MW ($p=0.812$), LogP ($p=0.233$), NumHDonors ($p=0.548$), NumHAcceptors ($p=0.423$). → Lack of significance mainly due to very small active group (n=2).

Fingerprint Calculation

- **Tool/Method used:** PubChem fingerprints
- **Number of bits/features generated:** 881
- **Why this method was selected:** PubChem fingerprints are fixed-length (881 bits), substructure-based, and widely used in QSAR studies for their good coverage of chemical patterns, interpretability, and compatibility with machine learning models. They capture presence/absence of functional groups and topological features effectively in small datasets like this one, making them suitable for subsequent modeling (e.g., regression or classification on pIC50 or activity class).