Exploratory data analysis and molecular descriptor calculations for FLT3 kinase inhibitors

1. Bioactivity Data Transformation

To ensure statistical robustness for machine learning, raw IC50 values were converted to the negative logarithmic scale (pIC50). The distribution is visualized using a histogram (fig.1) to ensure data quality and identify potential outliers.

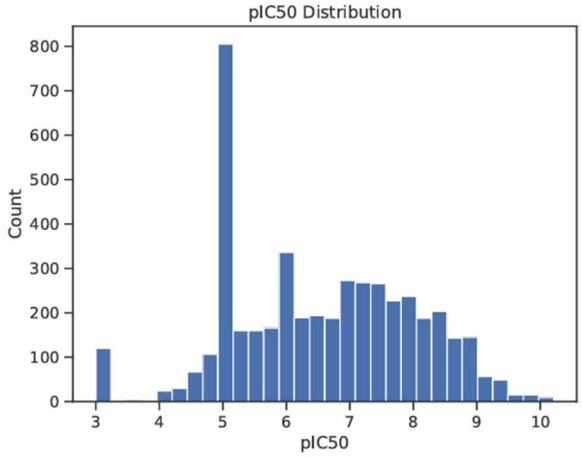

Figure 1. Distribution of pIC50 values across the dataset

The presence of outliers at the lower boundary (pIC50 = 3) corresponds to the 1,000,000 nM limit often found in ChEMBL, representing compounds with no measurable inhibitory activity at the highest tested concentration. These data points are essential as they provide the model with 'negative examples,' helping it to distinguish between pharmacologically relevant structures and inactive chemical space.

The distribution also features high-potency outliers with pIC50 values exceeding 9.0 (sub-nanomolar activity). While these represent a small fraction of the 4,647 compounds, they are the most pharmacologically significant molecules in the dataset. These outliers provide the 'positive signal' necessary for the model to identify the key structural features responsible for exceptional inhibitory potency.

The final dataset exhibits an **active-heavy distribution**, with active compounds outnumbering inactive ones in a ratio of approximately 1.7:1 (fig. 2).

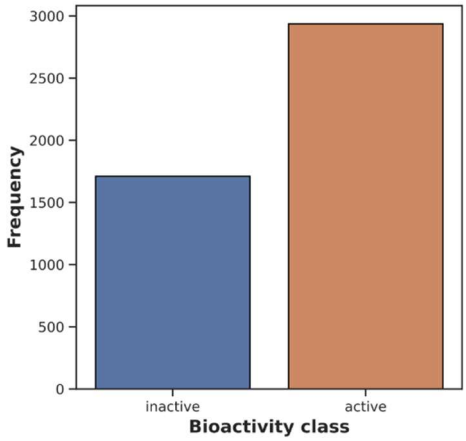

Figure 2. Frequency of active and inactive compounds in the dataset.

The overview below shows that the range of IC50 values covers pM to mM scale. The median is lower that 1000 nM threshold which again demonstrates the active-heavy dataset.

Log-transformation normalizes the data distribution, making it more symmetric and reducing the gap between the mean and median. This transformation successfully normalized the bioactivity distribution, reducing the impact of extreme outliers and compressing the data into a more manageable range (3.0 to 10.2) (table 1).

Table 1. Descriptive statistical summary of bioactivity values (standard_value in nM and log-transformed pIC50)

	standard_value	pIC50
count	4647.000000	4647.000000
mean	29807.677915	6.554864
std	158273.965204	1.466852
min	0.061000	3.000000
25%	19.255000	5.168002
50%	278.000000	6.555955
75%	6792.032500	7.715458
max	1000000.000000	10.214670

2. Physicochemical Space and Lipinski Validation

The drug-likeness of the FLT3 inhibitors was evaluated using Lipinski's descriptors.

A scatter plot analysis was conducted to visualize the chemical space distribution of the dataset based on Molecular Weight (MW) and LogP (fig.3). The distribution demonstrates that most active compounds and inactive compounds occupy a similar physicochemical space, primarily within the Lipinski-compliant boundaries. The overlapping clusters suggest that potency is driven by specific structural features rather than a simple global shift in molecular size or solubility.

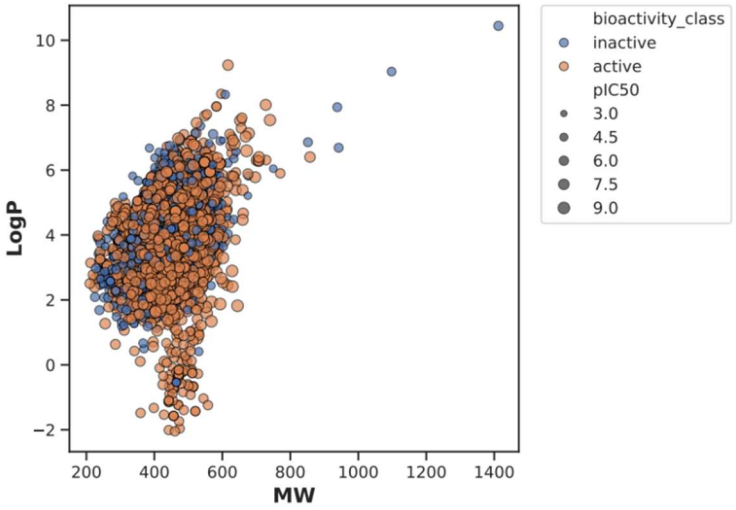

Figure 3. Scatter plot of Molecular Weight (MW) vs. LogP of both bioactivity classes

3. Statistical Significance (Mann-Whitney U Test)

A non-parametric Mann-Whitney U test was performed to validate the descriptors. All evaluated parameters, including Molecular Weight, LogP, and Hydrogen Bond Donors/Acceptors, yielded p-values significantly below the 0.05 threshold, confirming their predictive power for distinguishing compound activity (table 2).

Table 2. Results of the Mann-Whitney U test for physicochemical descriptors

Descriptor	Statistics	p	alpha	Interpretation
NumHAcceptors	2211996	5.87E-12	0.05	Different distribution (reject H0)
MW	2629346	0.007677	0.05	Different distribution (reject H0)
NumHDonors	4057871	2.88E-283	0.05	Different distribution (reject H0)
pIC50	5023496	0	0.05	Different distribution (reject H0)
LogP	2346974	0.000187	0.05	Different distribution (reject H0)

Differences between groups can also be presented as interquantile distributions (fig.4).

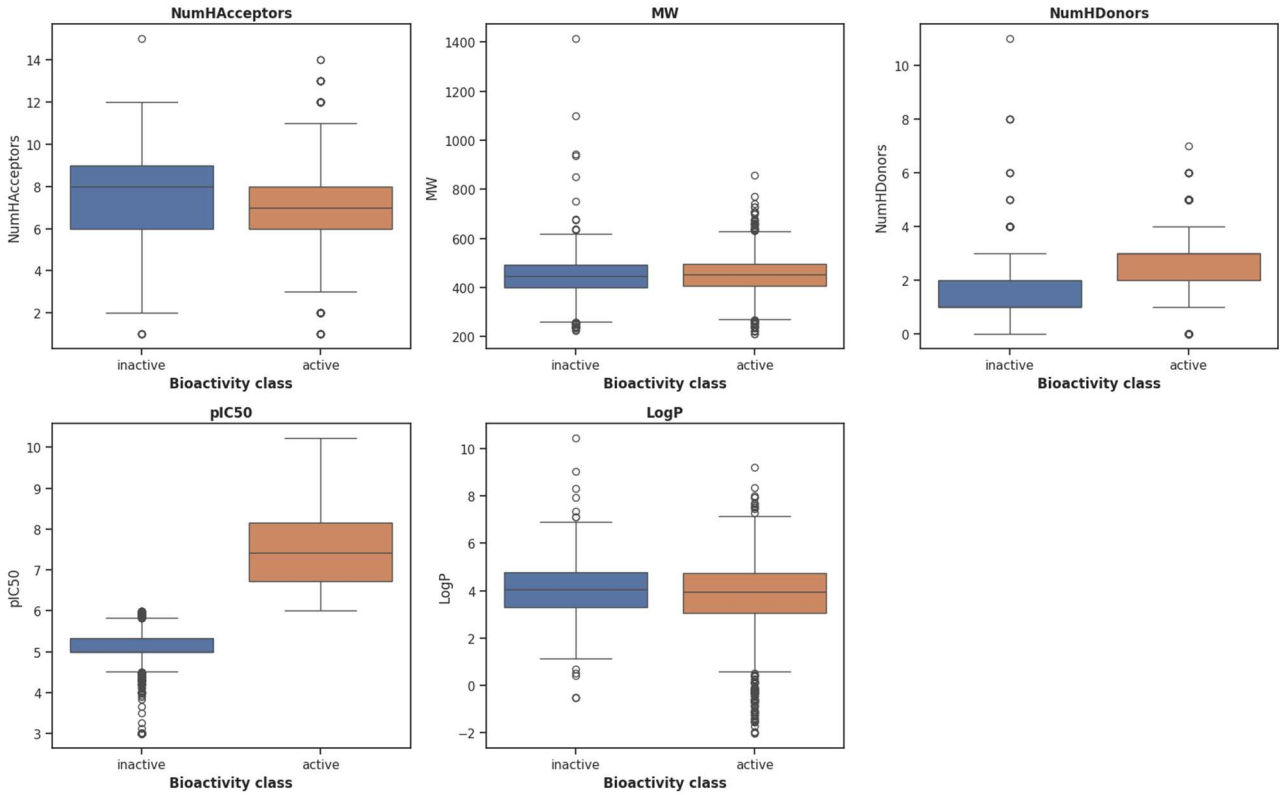

Figure 4. Boxplots of Lipinski descriptors and pIC50 categorized by activity

Conclusions based on statistical test and visualizations

All four Lipinski descriptors are statistically significant, as each yielded a p-value < 0.05 in the Mann-Whitney U-test. The number of hydrogen bond donors (NumHDonors) shows the highest significance, suggesting its critical role in molecular activity

No non-significant descriptors were identified among the calculated physicochemical properties in this dataset. However, Molecular Weight (MW) exhibited the highest p-value among the significant set, indicating a relatively weaker differentiating power compared to other features

NumHAcceptors

Active compounds tend to have slightly fewer hydrogen bond acceptors

Molecular Weight

The boxplot for Molecular Weight (MW) reveals that while both classes share a similar median, the active compounds are more concentrated within a tighter range, whereas the inactive compounds show a much broader distribution with significantly higher extreme outliers reaching up to 1400 Da.

NumHDonors

Active compounds generally have more hydrogen bond donors (typically 2 to 3) than inactive ones, which usually have only 1 or 2, as shown by the clear upward shift of the orange box. The lowest p-value indicates a strong relations between this descriptor and

IC50

The median of active class differs significantly from the inactive class median, which indicates robustness of the dataset. The outliers with extremely low values are those that were seen on the histogram.

LogP

For LogP, both active and inactive compounds share a similar central distribution, but the active set (orange) displays a significant cluster of low-lipophilicity outliers (negative LogP), which contrasts with the high-lipophilicity outliers seen in the inactive group.

4. Calculation of molecular descriptors via PaDEL-Descriptor

Chemical structures were successfully translated into high-dimensional numerical vectors using the PaDEL-Descriptor software. Two primary feature sets were generated: **topological 2D descriptors** (1,444 features) and **PubChem fingerprints** (881 bits).

5. Feature Selection and Noise Reduction

To optimize the dataset for future modeling, a **Variance Threshold** filter was applied. Constant features that provided zero information were identified and removed.

- The 2D descriptor set was refined to **1,209 informative variables**.
- The PubChem fingerprint set was refined to **599 unique informative bits**.