Introduction

Quantitative structure–activity relationship (QSAR) modelling, an approach that was introduced 60 years ago, is widely used in computer-aided drug design. In recent years, progress in artificial intelligence techniques, such as deep learning, the rapid growth of databases of molecules for virtual screening and dramatic improvements in computational power have supported the emergence of a new field of QSAR applications that we term ‘deep QSAR’.  (Tropsha et al., 2023)

The PPARs (peroxisome proliferator-activated receptors) have emerged as promising therapeutic targets among all nuclear receptors for developing novel pharmacotherapeutic candidates against insulin resistance, cancers, obesity, dyslipidemia, and cardiovascular disorders. The medicinal repertoire of PPAR*γ* antagonists spreads more wider than diabetic therapy in promoting osteoblast formation and depressing differentiation of adipose tissue. As a result, PPAR*γ* inhibitors can be well thought out to be potential aspirants for osteoporosis and obesity therapy. Additionally, PPAR*γ* inhibitors embody broad anticancer activity as well. For that reason, exploring PPAR*γ* inhibitors is of prodigious importance in the quest for novel drug candidates for pharmacotherapy of PPAR*γ*-associated metabolic disorders. (Ipsa Padhy et al., 2024)

The introduction could benefit from the following additions:

1. **Explanation of Deep QSAR**: Expand on the concept of "deep QSAR" introduced by Tropsha et al. (2023) and its significance in modern drug discovery. Provide a brief overview of how deep learning techniques have revolutionized QSAR modeling and its potential applications in virtual screening and drug design.
2. **Significance of PPARs as Therapeutic Targets**: Provide additional context on the significance of PPARs as therapeutic targets, emphasizing their role in various metabolic disorders and their potential for drug development. Highlight the breadth of therapeutic applications for PPARγ antagonists, including insulin resistance, obesity, osteoporosis, and cancer, as discussed by Ipsa Padhy et al. (2024).
3. **Rationale for Studying PPARγ Inhibitors**: Discuss the rationale behind focusing on PPARγ inhibitors for drug development, considering their broad therapeutic potential and their importance in addressing PPARγ-associated metabolic disorders. Emphasize the need for novel drug candidates in pharmacotherapy, particularly in light of the growing prevalence of metabolic disorders worldwide.

* PPAR gamma is a nuclear receptor involved in regulating fat and carbohydrate metabolism.
* Researchers are interested in finding compounds that can activate (agonize) PPAR gamma, potentially for applications in treating diabetes and other metabolic disorders.

Methodology

Quantitative Structure-Activity Relationship (QSAR) modeling is outlined. Initially, data collection and preprocessing are performed by selecting the target protein Peroxisome proliferated activated receptor gamma (PPARγ) associated with diabetes from the ChEMBL database and retrieving bioactivity data reported as IC50 values in nM units. Missing data is handled by removing compounds with incomplete information.

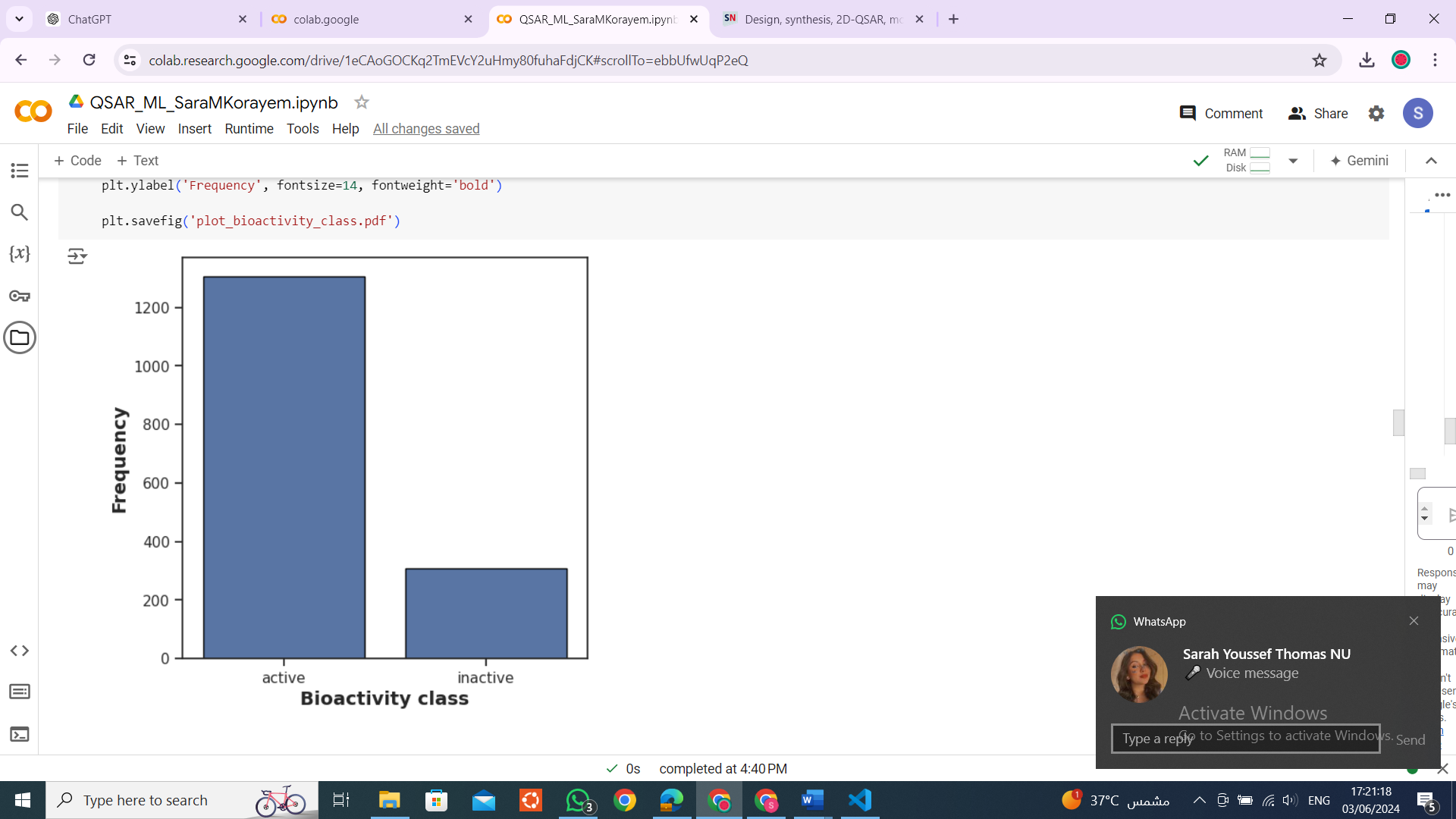
Following data exploration, feature engineering involves calculating molecular descriptors and selecting relevant features for model development. Bioactivity data was classified into "active," "inactive," or "intermediate" categories based on IC50 values, facilitating subsequent analysis. Relevant features, including molecule identifiers, canonical SMILES representing chemical structures, and bioactivity values, are selected and combined with corresponding bioactivity classes. Lipinski descriptors, pivotal for evaluating drug-likeness according to Lipinski's Rule of Five, are then computed from canonical SMILES. These descriptors encompass molecular weight, LogP, and counts of hydrogen bond donors and acceptors. Finally, the calculated descriptors are merged with the original dataset, forming a comprehensive DataFrame for further exploration and QSAR modeling.

IC50 values are converted to a negative logarithmic scale, termed as pIC50, ensuring uniformity in bioactivity data across different scales. This is achieved through a custom function, pIC50(), which transforms IC50 values from nanomolar to molar units, applies the negative logarithmic transformation, and integrates the results into a new DataFrame column while removing the original standard\_value\_norm column. Subsequently, the normalization of standard values is carried out to cap values exceeding 100,000,000 nM and prevent negative logarithmic values. Relevant columns are retained for further analysis, and compounds labeled as 'intermediate' in the bioactivity class are excluded.

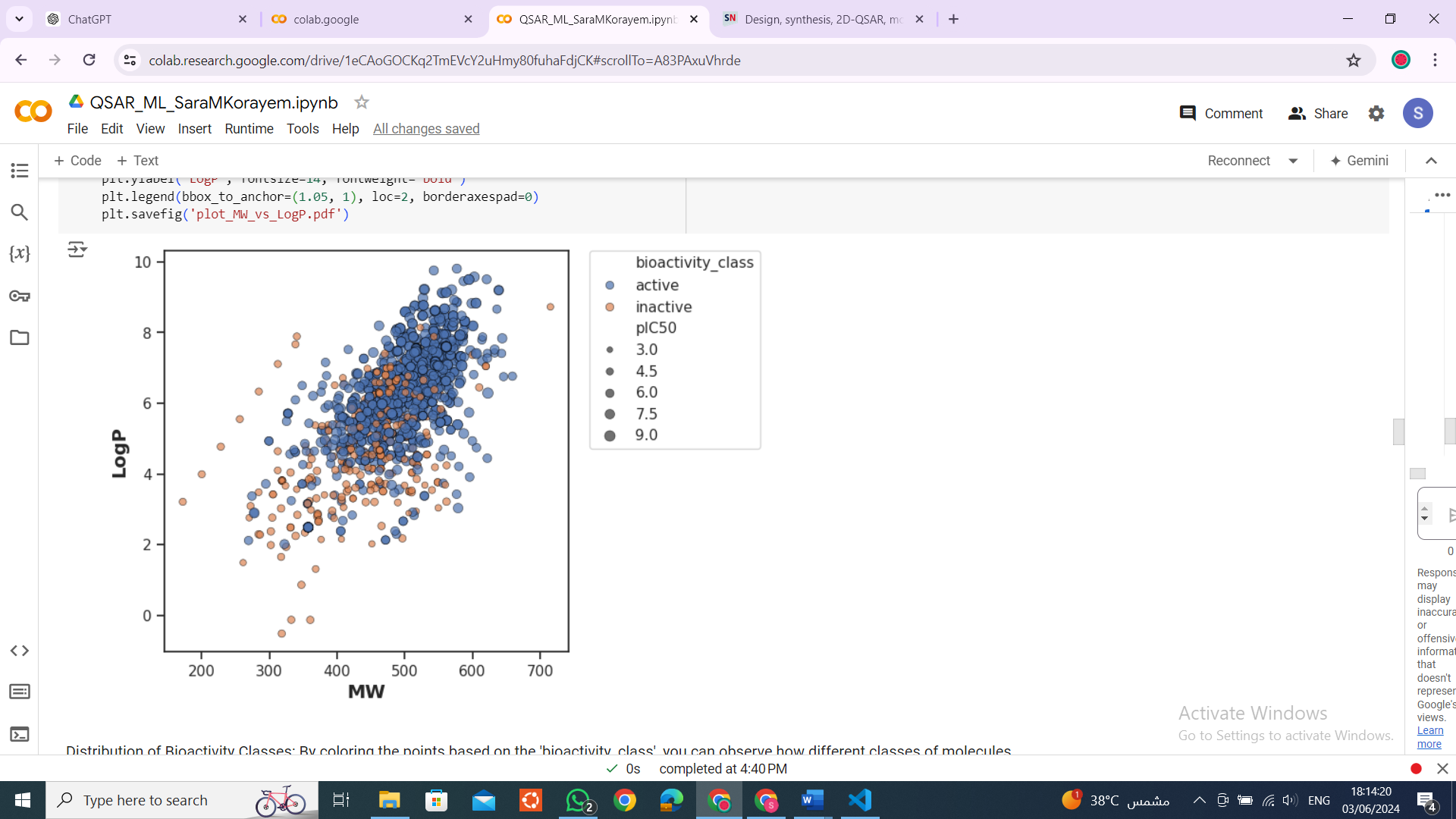
Exploratory Data Analysis (EDA) follows, employing Lipinski descriptors to explore the chemical space occupied by 'active' and 'inactive' compounds. Visualizations such as scatter plots and box plots are generated to visualize the distribution of bioactivity classes and assess the significance of differences in molecular descriptors between classes. Statistical tests, including the Mann-Whitney U test, are conducted to evaluate these differences. Results are interpreted to discern meaningful relationships and identify descriptors crucial for distinguishing between bioactivity classes. Finally, processed data and visualizations are exported for further analysis and documentation.

Dataset was prepared for building a regression model of inhibitors using the random forest algorithm. Initially, molecular descriptors are calculated using the PaDEL-Descriptor tool, enabling the quantitative description of compounds in the dataset. The bioactivity data, containing pIC50 values, is loaded from a previously pre-processed ChEMBL dataset. The dataset is then divided into input features (X) and the target variable (Y), followed by the removal of low variance features. Subsequently, the data is split into training and testing sets in an 80/20 ratio. A regression model is built using the Random Forest algorithm, with 100 estimators. Model performance is evaluated using R-squared values and Root Mean Squared Error (RMSE), and a scatter plot of experimental versus predicted pIC50 values is generated. Additionally, the code showcases the use of the LazyPredict library to compare multiple machine learning algorithms for regression modeling. The top-performing model, which was Random Forest is selected based on adjusted R-squared, R-squared, RMSE, and computation time metrics. Finally, the selected model is trained and saved for deployment, with label encoding applied to resolve potential issues with the 'molecule\_chembl\_id', ‘canonical\_smiles’ and ‘bioactivity\_class’ columns. The process concludes with hyperparameter tuning using GridSearchCV and saving the trained model using both joblib and pickle libraries. The saved model was used for deployment using streamlit.

Results

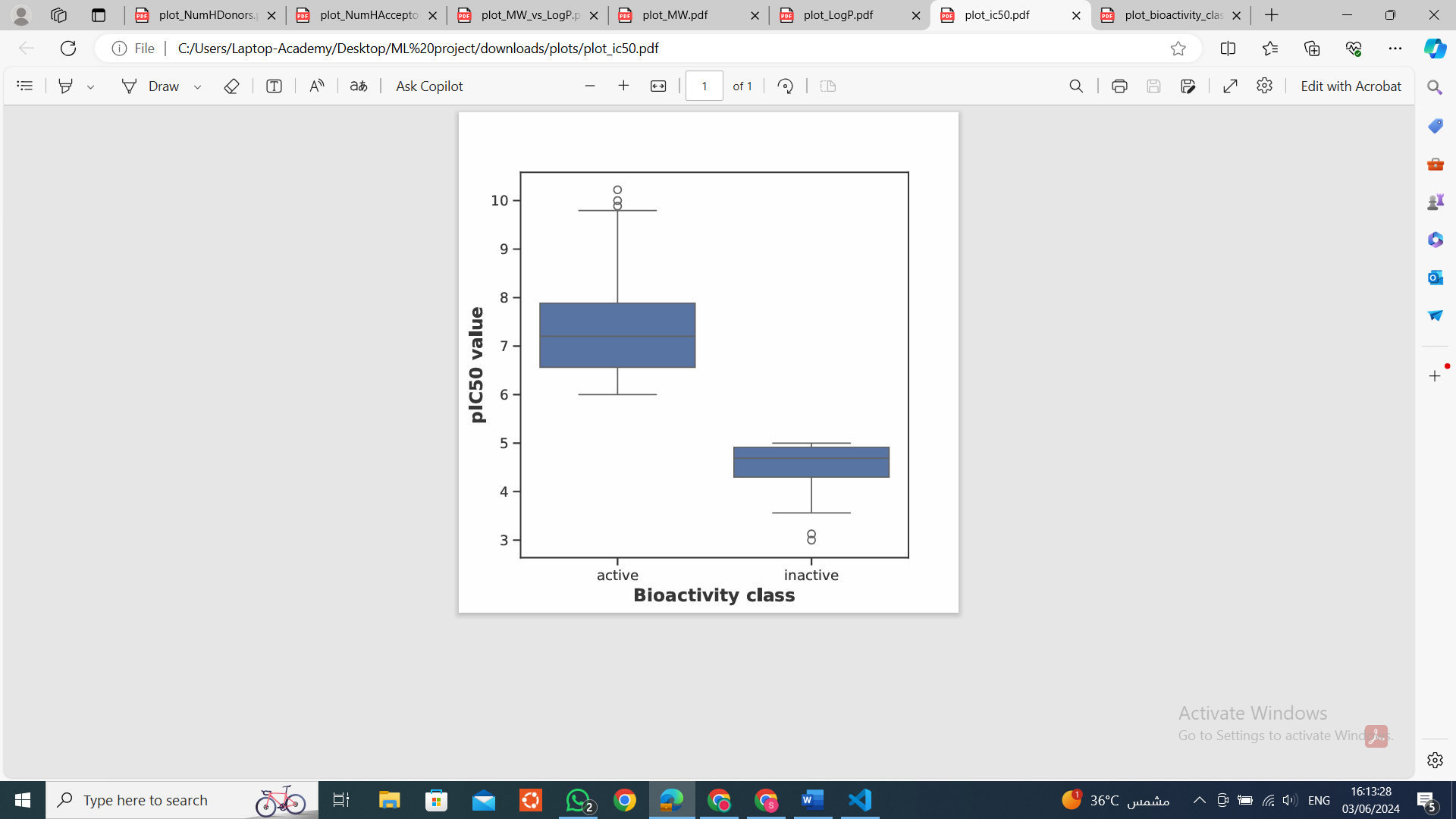


The bar chart illustrates the distribution of compounds across different bioactivity classes. The x-axis of the chart represents the bioactivity class, with two categories: "active" and "inactive." Meanwhile, the y-axis denotes the frequency, ranging from 0 to 1200, indicating the number of compounds classified into each category. This observation underscores a notable prevalence of compounds classified as "active" within the dataset, suggesting a higher frequency of bioactive compounds compared to inactive ones.



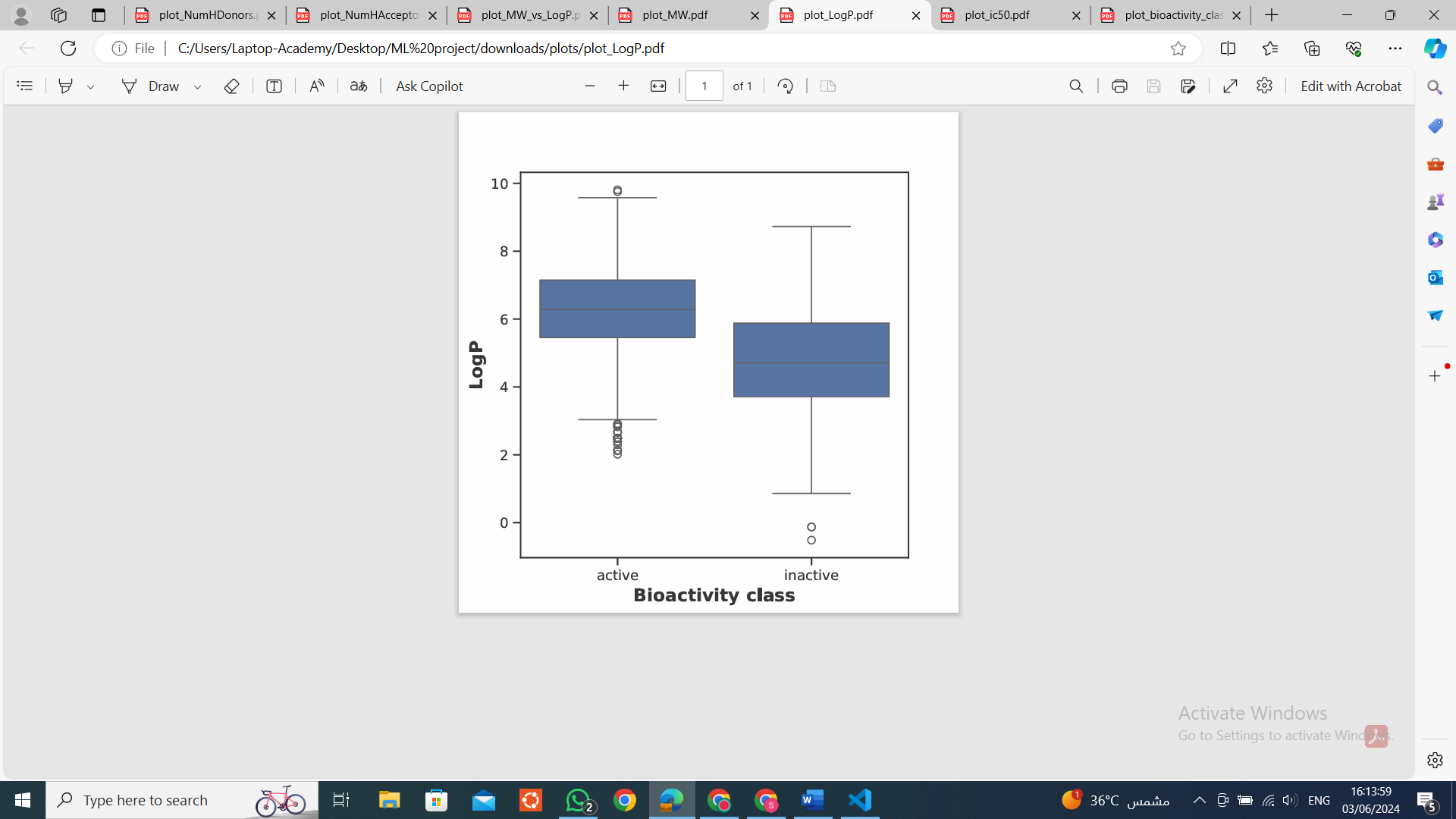
The analysis of molecular weight (MW) and LogP (the logarithm of the octanol-water partition coefficient) reveals valuable insights into the molecular characteristics influencing the activation of PPAR gamma. A lower molecular weight suggests a smaller molecule size, and the observed trend implies that smaller molecules might exhibit greater effectiveness in activating PPAR gamma. Conversely, LogP indicates the hydrophobicity of a molecule, with lower LogP values indicating higher hydrophilicity. The analysis suggests a trend where more hydrophilic molecules, characterized by lower LogP values, may possess greater potential for activating PPAR gamma. These findings provide valuable guidance for further exploration of structure-activity relationships and the rational design of compounds with enhanced PPAR gamma activation properties.

The Mann-Whitney U test results using Lipinski descriptors



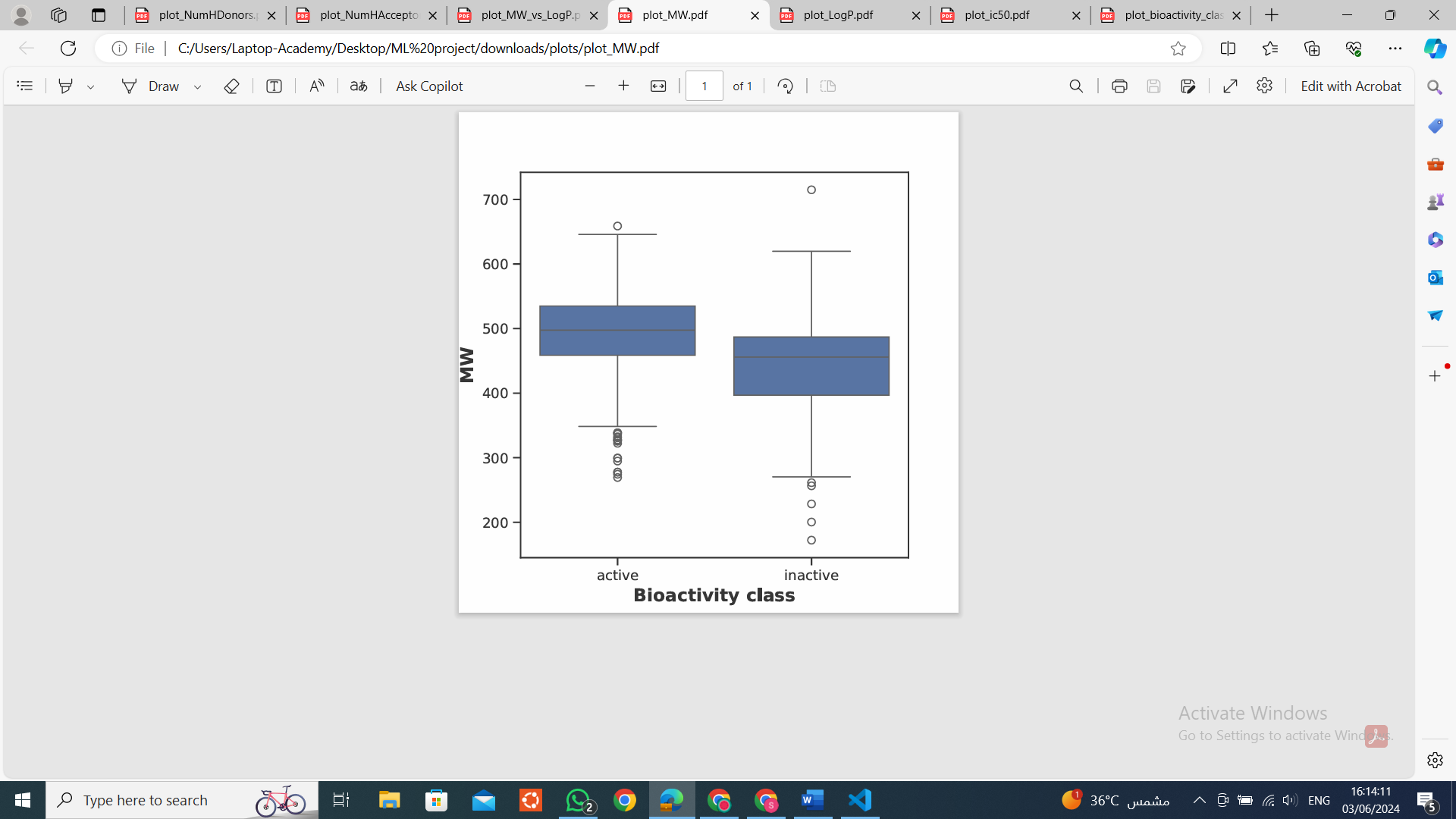
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Descriptor** | **Statistics** | **p** | **alpha** | **Interpretation** |
| pIC50 | 404860.0 | 2.154229e-165 | 0.05 | Different distribution (reject H0) |

The Mann-Whitney U test was conducted to compare the distribution of pIC50 values between active and inactive compounds. The test yielded a statistic of 404860.0 and a p-value of 2.154229e-165. With a significance level (alpha) set at 0.05, the obtained p-value was significantly lower, indicating a statistically significant difference in the distribution of pIC50 values between active and inactive compounds. Consequently, we reject the null hypothesis (H0) and conclude that there is a different distribution of pIC50 values between active and inactive compounds. This finding provides valuable insight into the molecular characteristics associated with biological activity and guides further exploration of structure-activity relationships, ultimately informing drug discovery efforts.



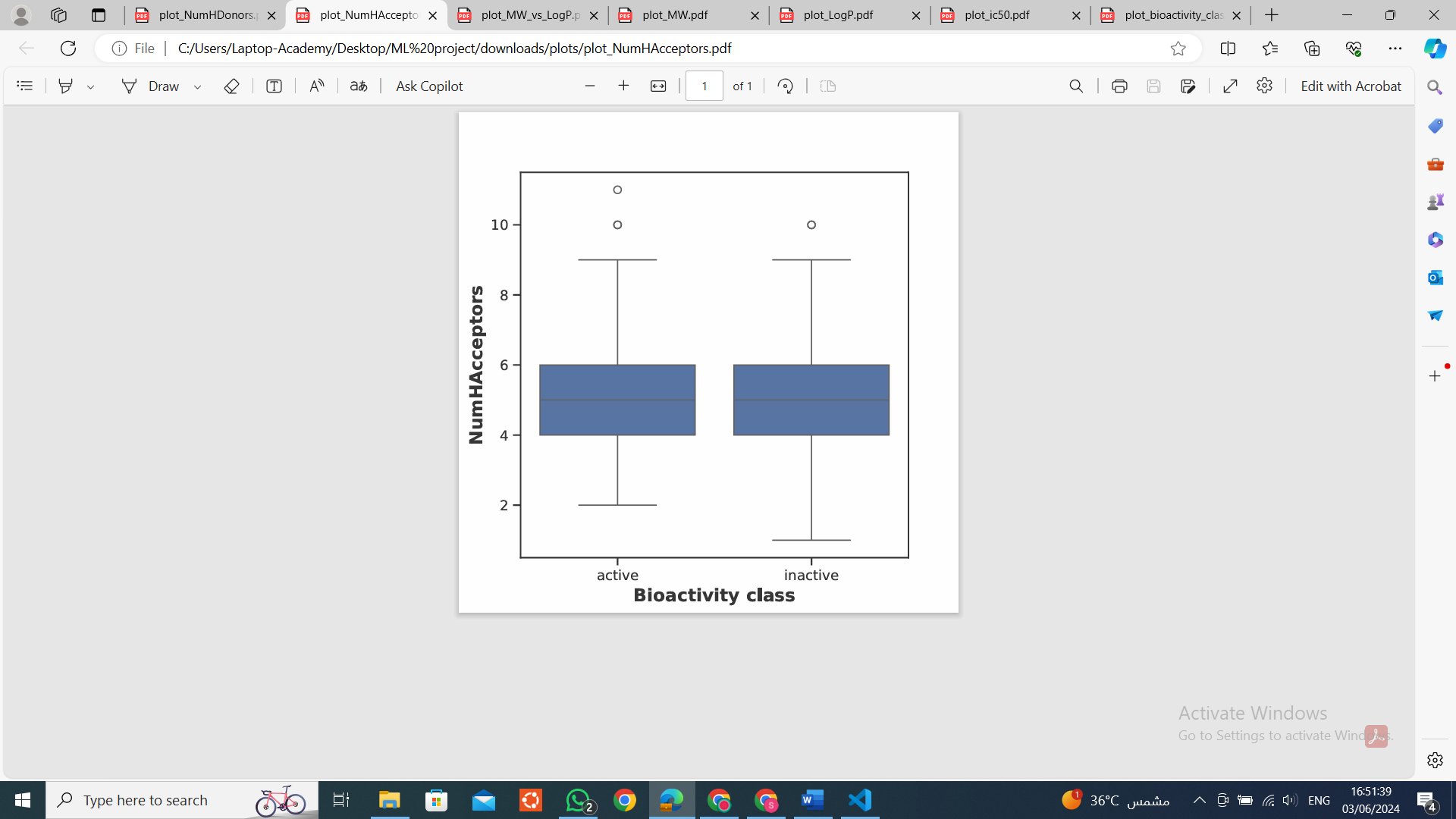
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Descriptor** | **Statistics** | **p** | **alpha** | **Interpretation** |
| LogP | 307905.0 | 2.902892e-46 | 0.05 | Different distribution (reject H0) |

The Mann-Whitney U test was conducted to compare the distribution of LogP values between compounds classified as active and inactive. The test yielded a statistic of 307905.0 and a p-value of 2.902892e-46. With a significance level (alpha) set at 0.05, the obtained p-value was significantly lower, indicating a statistically significant difference in the distribution of LogP values between active and inactive compounds. Therefore, we reject the null hypothesis (H0) and conclude that there is a different distribution of LogP values between active and inactive compounds. This finding suggests that LogP, a measure of lipophilicity, is a distinguishing factor between active and inactive compounds, providing valuable insight into the molecular characteristics associated with biological activity.



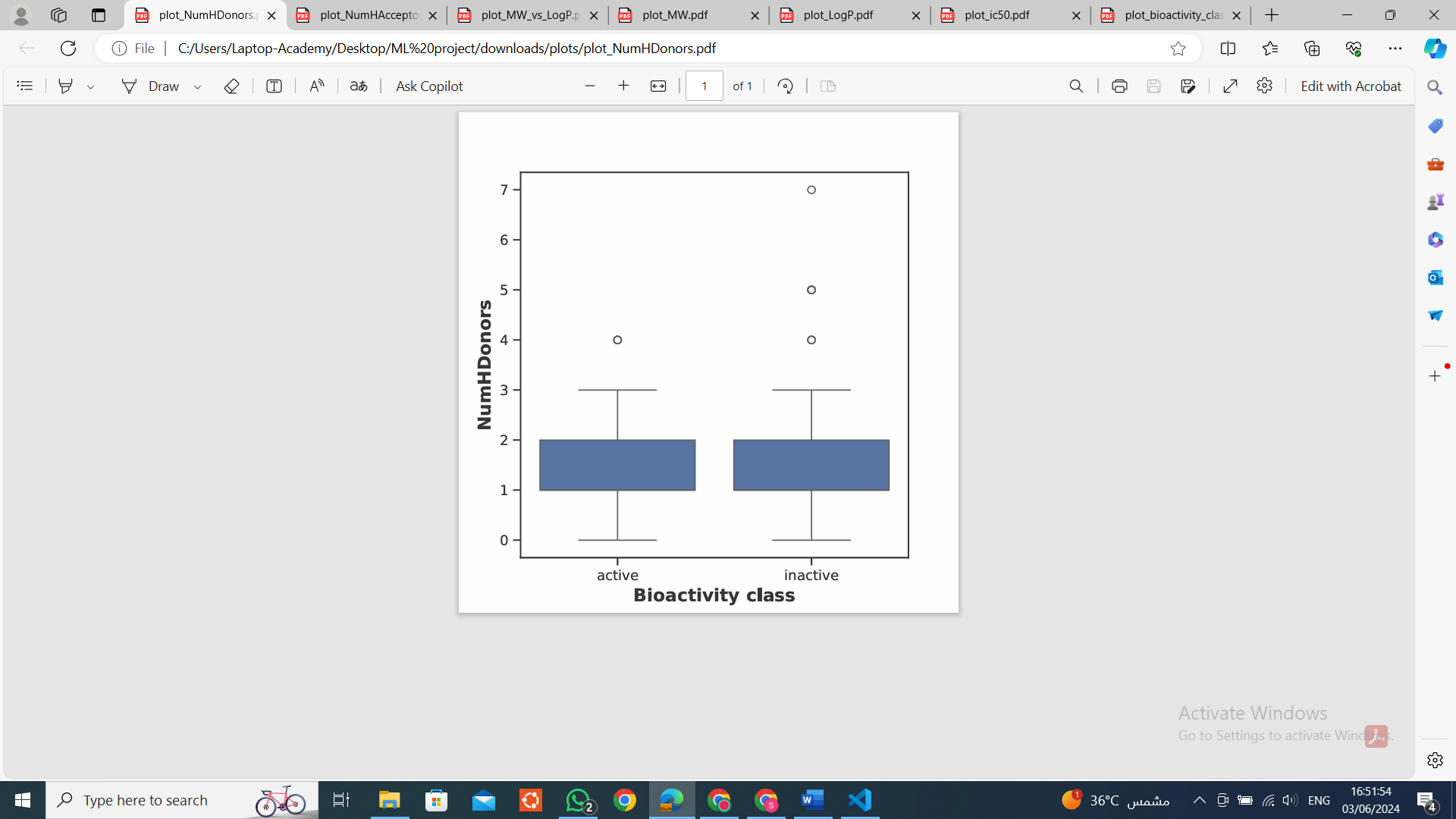
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Descriptor** | **Statistics** | **p** | **alpha** | **Interpretation** |
| MW | 291456.0 | 1.863988e-33 | 0.05 | Different distribution (reject H0) |

The Mann-Whitney U test was conducted to compare the distribution of molecular weight (MW) values between compounds classified as active and inactive. The test yielded a statistic of 291456.0 and a p-value of 1.863988e-33. With a significance level (alpha) set at 0.05, the obtained p-value was significantly lower, indicating a statistically significant difference in the distribution of MW values between active and inactive compounds. Therefore, we reject the null hypothesis (H0) and conclude that there is a different distribution of MW values between active and inactive compounds. This finding suggests that molecular weight is a distinguishing factor between active and inactive compounds, providing valuable insight into the molecular characteristics associated with biological activity.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Descriptor** | **Statistics** | **p** | **alpha** | **Interpretation** |
| NumHAcceptors | 176441.5 | 0.000321 | 0.05 | Different distribution (reject H0) |

The Mann-Whitney U test was conducted to compare the distribution of the number of hydrogen bond acceptors (NumHAcceptors) between compounds classified as active and inactive. The test yielded a statistic of 176441.5 and a p-value of 0.000321. With a significance level (alpha) set at 0.05, the obtained p-value was significantly lower than the alpha level, indicating a statistically significant difference in the distribution of NumHAcceptors between active and inactive compounds. Therefore, we reject the null hypothesis (H0) and conclude that there is a different distribution of the number of hydrogen bond acceptors between active and inactive compounds. This finding suggests that the number of hydrogen bond acceptors is a distinguishing factor between active and inactive compounds, providing valuable insight into the molecular characteristics associated with biological activity.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Descriptor** | **Statistics** | **p** | **alpha** | **Interpretation** |
| NumHDonors | 220837.0 | 0.004085 | 0.05 | Different distribution (reject H0) |

The Mann-Whitney U test was conducted to compare the distribution of the number of hydrogen bond donors (NumHDonors) between compounds classified as active and inactive. The test yielded a statistic of 220837.0 and a p-value of 0.004085. With a significance level (alpha) set at 0.05, the obtained p-value was lower than the alpha level, indicating a statistically significant difference in the distribution of NumHDonors between active and inactive compounds. Therefore, we reject the null hypothesis (H0) and conclude that there is a different distribution of the number of hydrogen bond donors between active and inactive compounds. This finding suggests that the number of hydrogen bond donors is a distinguishing factor between active and inactive compounds, providing valuable insight into the molecular characteristics associated with biological activity.

Discussion

key findings gives insights about the distribution of compounds across different bioactivity classes and the molecular attributes influencing PPAR gamma activation. The analysis of bioactivity class distribution revealed a significant prevalence of "active" compounds compared to "inactive" ones, underscoring the dataset's composition and its relevance to drug discovery endeavors. Furthermore, our examination of molecular weight (MW) and LogP shed light on their roles in PPAR gamma activation, suggesting that smaller molecule sizes and higher hydrophilicity may enhance effectiveness in activating PPAR gamma. Additionally, the Mann-Whitney U test results for Lipinski descriptors unveiled significant differences between active and inactive compounds across various parameters such as pIC50 values, LogP, MW, NumHAcceptors, and NumHDonors. These findings deepen our understanding of structure-activity relationships and provide valuable insights into the molecular characteristics associated with biological activity, offering a roadmap for informed drug design strategies targeting PPAR gamma activation in diabetes treatment.

Ipsa Padhy, Banerjee, B., Ganga, P., Gupta, P. P., & Sharma, T. (2024). Design, synthesis, 2D-QSAR, molecular dynamic simulation, and biological evaluation of topiramate–phenolic acid conjugates as PPARγ inhibitors. *Future Journal of Pharmaceutical Sciences*, *10*(1). https://doi.org/10.1186/s43094-024-00617-1

Tropsha, A., Isayev, O., Varnek, A., Schneider, G., & Cherkasov, A. (2023). Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. *Nature Reviews Drug Discovery*, 1–15. https://doi.org/10.1038/s41573-023-00832-0