Neurokinin-1 Report

Akanna Chukwumerije[[1]](#footnote-0), Leah Joseph[[2]](#footnote-1), Courage Idahosa[[3]](#footnote-2), Kamui Albright[[4]](#footnote-3)

## Abstract

Our project relates to drug discovery in the modern world. Specifically, we analyze the results of the Neurokinin-1 (NK1) dataset. The pharmaceutical industry uses quantitative structure-activity relationships (QSAR) for prediction, leading to further prioritization and discovery. Through our analysis, pharmaceuticals can substantially reduce the experimental stage. Our model compares the target variable (ACT) with the D variable (feature) to recognize patterns and relationships within the individual molecule. We used many different models to conduct our project, including random forest, K-nearest neighbors, support vector regression (SVR), and neural networks. Our random forest model displayed the highest R value, indicating a good fit as the model can capture a significant portion of the variability in the dependent variable, making its predictions more reliable. In addition, the random forest model illustrated the lowest mean squared error value (MSE) and mean absolute error (MAE). In contrast, our neural network model represents a poorly performing model with the lowest R ^2 and the highest MSE.

## Introduction

Artificial Intelligence has revolutionized the pharmaceutical industry. The integration of AI and medicine has impacted a variety of pharmaceutical sectors. We will utilize datasets to analyze drug discovery. Antiemetic Neurokinin-1 (NK-1) receptor antagonists are a type of antimetic drug. The discovery of NK-1 is vital in chemotherapy research, particularly in suppressing nausea and vomiting after patients receive chemotherapy (Ibrahim). The figure below demonstrates an NK-1 receptor antagonist called Aprepitant, the first NK-1 receptor antagonist approved as an anti-nausea for chemotherapy.

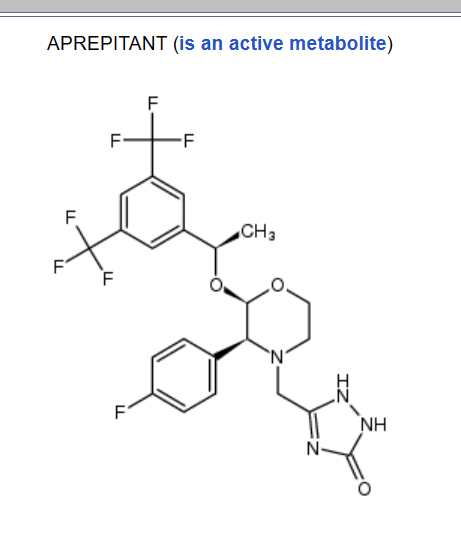


Figure 1: Molecular illustration of NK1 Receptor Aprepitant

Through our project, we analyze the relationships between NK-1 structures and target molecules. The data takes 15 target molecules and over 10,00 compounds per target molecule. The connection between the target and compound molecule is essential for prediction. Our approach is to recognize the receptor's and target molecules' molecular structure and analyze their relationship along with the relationships between the compounds. For our project, we utilized a variety of models, such as random forest, K nearest neighbors (KNN), support vector regression (SVR), and neural networks. When comparing the regression models, we noticed that our neural network model had more outliers but was narrower. We also did a residual comparison of our models to illustrate. The random forest model exhibited the highest R^2 value, suggesting a good fit as it effectively captures the dependent variable's variability, enhancing its predictions' reliability. The random forest model demonstrated the lowest values for MSE and MAE. The KNN and SVM models performed relatively similarly with nearly identical R^2, MSE, and MAE. In contrast, the neural network model performed less optimally with the lowest R^2 and the highest MSE among the models.

## Background

Drug discovery is a complex and resource-intensive process within the pharmaceutical industry, and technological advancements have played a pivotal role in streamlining this endeavor. One prominent approach in modern drug discovery is using quantitative structure-activity relationships (QSAR), a method that establishes correlations between the chemical structure of a compound and its biological activity. This approach allows researchers to predict the potential efficacy of a drug candidate, facilitating further prioritization and expediting the drug discovery pipeline.

In the context of our project, we focus on analyzing the Neurokinin-1 (NK1) dataset. The NK1 receptor is a crucial target in neuropharmacology, with implications for various therapeutic areas, including pain management and psychiatric disorders. Analyzing this dataset provides a wealth of information about the molecular interactions and activities associated with compounds targeting the NK1 receptor. Our specific methodology involves comparing the target variable (ACT) with the D variable (feature) within the individual molecules in the NK1 dataset. The target variable typically represents the biological activity or efficacy of the compounds, while the D variable encompasses various molecular features. By establishing patterns and relationships between these variables, our model aims to recognize critical attributes contributing to the molecules' effectiveness in targeting the NK1 receptor. The significance of this approach lies in its potential to reduce the experimental stage of drug development substantially. By leveraging the insights gained from the analysis of the NK1 dataset, pharmaceutical researchers can make more informed decisions about which compounds are likely to exhibit desired pharmacological effects. This accelerates the drug discovery process and minimizes the need for extensive experimental testing, saving time and resources. Several related studies have explored similar methodologies in QSAR and drug discovery. Researchers have applied machine learning techniques to diverse datasets, aiming to uncover structure-activity relationships and predict the biological effects of novel compounds. These studies contribute to the growing body of knowledge surrounding computational approaches to drug discovery, emphasizing the importance of leveraging advanced analytics to enhance decision-making in identifying potential drug candidates. In summary, our project stands at the intersection of modern drug discovery, QSAR methodology, and the analysis of the NK1 dataset. By employing advanced computational techniques, we aim to advance our understanding of the molecular basis for drug efficacy, ultimately contributing to the more efficient and effective development of pharmaceutical compounds targeting the NK1 receptor.

## Data

The data I'm working with for my project comprises information on chemical structures represented by molecular identifiers (M\_18, M\_225, M\_369, M\_372, M\_408, M\_579, M\_587, etc.) along with associated numerical values for a property denoted as 'Act'. Additionally, the dataset includes columns labeled as descriptors (D\_2, D\_3, D\_4, D\_36, D\_37, D\_38, D\_39, D\_40, D\_41, D\_42, D\_43, D\_44, D\_45, D\_46, D\_48, D\_49, D\_50), likely containing various features derived from these chemical structures. This data is acquired from a source that provides chemical property information or molecular characteristics. I am working with a specific subset of this data, comprising a few rows (M\_18, M\_225, M\_369, M\_372, M\_408, M\_579, M\_587, etc.) and several descriptor columns. As for preprocessing, our training and testing datasets were changed to ensure harmonization with the descriptors' columns. This preprocessing step aimed at achieving alignment between the available data and the specific descriptors needed for our modeling task. During this process, a notable observation emerged—a section within the dataset comprised rows where all descriptor columns contained zero values. These zero values do not indicate missing data but represent meaningful chemical features.

## Methods

Our methods for this experiment went for a comprehensive approach leveraging concepts and skills acquired during this semester. Firstly, we recognized the task as a regression problem, requiring the prediction of continuous values. To address this, we employed various regression algorithms, including Random Forest, K-Nearest Neighbors, Support Vector Machine (SVM), and a Neural Network built using TensorFlow/Keras. We began by loading and preparing the data, splitting it into distinct training and test sets, adhering to best practices to avoid overfitting and ensure robust model evaluation. Each model was meticulously constructed, fitted to the training data, and evaluated using Mean Squared Error (MSE) on the held-out test data. This approach was chosen for its inclusivity, allowing us to explore diverse models to capture potential nuances in the dataset.

Moreover, it facilitated a holistic comparison of these models, considering their unique strengths and weaknesses. This methodology aligns with the semester's teachings, emphasizing the importance of model selection, evaluation, and applying diverse algorithms to discern the most suitable approach for a specific problem. The aim was to leverage various regression techniques and empirical evaluation to ascertain the model that best generalizes and performs optimally on the given dataset.

## Experiments

After training each model, the predicted target variables were compared to the target variables in the test dataset to evaluate their accuracy. According to the scatter plot, all models showed a diagonal trend with a similar density, with the neural network model having results slightly closer to the diagonal line, the random forest being the second, SVM being the third, and KNN being the last. Thus, all those models predicted the target variable with a noticeable accuracy. The density increases gradually as the value increases, so it indicates that the models are more efficient in predicting molecular activities of more bioactive compounds; however, for molecules that have 6 for activity value. NN showed numerous predictions that didn’t follow an appropriate trend, and a few outliers are also present. This error-indicative result can be seen more clearly on the residual comparison graph. Additionally, another notable result was found in the error metrics comparison. Here, the RF model was shown to be the clear winner among them, with lower MSE and MAE and higher R2 values. While KNN and SVM models had similar results, the neural network scored significantly worse with an extremely high MSE value and low R2 value. These results support the reasoning for the trend of the random forest model and SVM being the most popular and conventional models for molecular bioactivity predictions. One fact to note here, though, is that NN was showing some excellent predictions, so there is a potential for neural networks to be used in the future. However, it would require more precise and careful analysis and modification of datasets.

## Conclusion

In conclusion, evaluating various predictive models for molecular bioactivity, particularly concerning NK1 drugs, indicates insights into the potential effectiveness or challenges in predicting the bioactivity of compounds targeting the NK1 receptor.

The performance analysis suggests that specific models, notably Random Forest and Support Vector Machines, exhibit stronger predictive capabilities in assessing the bioactivity of compounds related to NK1 receptors. These models showcase lower errors (MSE and MAE) and higher R-squared values, implying a better ability to predict the effectiveness or potency of drugs targeting the NK1 receptor. Conversely, while Neural Networks showed sporadic instances of accurate predictions, their overall performance in this context was less favorable, marked by higher MSE and lower R-squared values. This suggests that while Neural Networks possess potential, their application for predicting bioactivity related to NK1 drugs might require more extensive data refinement, model tuning, or specific modifications to enhance accuracy. Moving forward, to help the potential of modeling in drug discovery targeting the NK1 receptor, expanding the scope of molecular research in improving how specific drugs can work toward targeting and studying the NK1 receptor. Currently, NK1 is used for pain modulation and mood regulation, exploring its role in gastrointestinal disorders, mental health treatments, and neurological conditions like Alzheimer's. Combining NK1 receptor modulation with existing therapies, personalized medical treatment, advancements in drug delivery, and investigation of natural compounds offer promising paths for future medical innovations. Continued research with AI will transform the medical industry.

## 

## References

*Browse E-DRUG3D Database*, chemoinfo.ipmc.cnrs.fr/MOLDB/browse.php?query=915\_1434. Accessed 9 Dec. 2023.

*Deep Neural Nets as a Method for Quantitative Structure–Activity...*, pubs.acs.org/doi/10.1021/ci500747n. Accessed 10 Dec. 2023.

Ibrahim, Michael A., and Charles V Preuss. “Antiemetic Neurokinin-1 Receptor Blockers - Statpearls - NCBI Bookshelf.” *Antiemetic Neurokinin-1 Receptor Blockers*, 24 Mar. 2023, www.ncbi.nlm.nih.gov/books/NBK470394/.

Vivek. “QSAR-Data for Drug Discovery.” *Kaggle*, 12 Oct. 2022, www.kaggle.com/datasets/vivek153/data-root/data?select=NK1\_training\_disguised.csv.

“Google Colaboratory.” *Google Colab*, Google, colab.research.google.com/drive/17pxGT6fFkNA6C-uEKM8fN3NNMYrud\_Q3?usp=sharing. Accessed 11 Dec. 2023.

1. Data, Methods, and Conclusion. [↑](#footnote-ref-0)
2. Abstract and Introduction. Group discussion and coding. 25% contribution [↑](#footnote-ref-1)
3. Background. [↑](#footnote-ref-2)
4. Experiments. [↑](#footnote-ref-3)