HackBio Internship Program

February 2025

Stage 3 Project Report

Drug Discovery

Favour Imoniye, Meltem Kutnu

Abstract

Adenosine deaminase (ADA2) is an essential enzyme in purine metabolism that helps maintain the immune system, cell proliferation and differentiation. Changes in ADA2 are associated with various immune system-related conditions, therefore targeting ADA2 with small molecules can aid in the development of new compounds against this severe condition. In this project, a chemical descriptor matrix of more than 10000 compounds was analyzed using regression and dimensionality reduction techniques. Regression analyses showed increased variations in the most important chemical features in determining the docking scores, pointing to issues in pre-processing and the random sampling of the data, as well as the variability in the features themselves. Pre- and post-feature selection resulted in altered statistical values in random forest regression as a possible result of removing variables with constant values or the choice of parameter values used in recursive feature elimination (RFE). Additionally, using the same RFE control for both regression models could have decreased the accuracy of the models. The results of the PCA visualisation yielded a compressed structure of data points that were highly concentrated at the center, signaling the minimal amount of variance captured by the first two principal components. Contrastingly, UMAP generated a globally representative depiction of the complex structures present within chemical space. The distinction in the results may be attributed to their operational mechanisms, with PCA relying on the presence of linear relationships, while UMAP is influenced by the non-linear relationships present in the data.

1. Introduction

As part of the Stage 3 task of the HackBio Internship Program in February 2025, a chemical descriptor matrix was analyzed. Chemical descriptor matrices include diverse structural properties of small molecules and are essential for ligand-based similarity searching (Grisoni et al., 2016). They consist of active ligands and decoy ligands that are not active against the target protein, but have similar properties to the target molecule. The matrix of interest for this project comprises over 10000 compounds docked against adenosine deaminase (ADA2). ADA2 is a key enzyme in purine metabolism, contributing to various cellular responses by the degradation of extracellular adenosine. ADA2's primary function in humans is the development and maintenance of the immune system, but it may also play a role in the regulation of cell proliferation and differentiation, independent from its enzymatic activity (Zavlaviov et al., 2010).

Variations in ADA2 are associated with VAIHS (vasculitis, autoinflammation, immunodeficiency and hematologic defects syndrome) and Sneddon syndrome. Additionally, increased levels of ADA2 were found to be correlated with rheumatoid arthritis, psoriasis and sarcoidosis. Therefore, identification of small molecules potentially effective against ADA2 can facilitate the development of novel compounds against this severe condition.

In light of the information above, the following questions were attempted to be answered in this project:

- How can PCA and UMAP models disentangle the structural features present in the dataset?
- Can we visibly differentiate decoys from the active molecules?
- Can we effectively predict the docking score using the chemical features of the ligands alone?
- Which chemical features can be used to determine if a ligand binds to a target protein?

This project aims to represent the chemical space using linear and non-linear dimensionality reduction methods to visualise the different ways in which they organise the global and local structure of the original data. The same information is employed for linear regression and random forest regression analyses to verify if we can utilize the structural properties of ligands alone for the prediction of docking scores, and if so, which features are relevant.

2. Methods

2.1. Principal Component Analysis (PCA) and Uniform Manifold Approximation (UMAP)

Principal component analysis (PCA) is a reduction technique that is used to transform high-dimensional data into a lower dimensional representation which captures a vast proportion of the variations and patterns in the original data. It is widely adapted for exploratory data analysis due to its ability to simplify complex datasets with multiple variables and eliminate multicollinearity. Uniform manifold approximation (UMAP), on the other hand, is a recently developed dimension reduction technique that uses Riemannian manifolds to examine the complex and non-linear relationships present within a dataset. It is notable for its scalability and its distinctive tendency to preserve a balanced local and global structure of the data compared to other dimensionality reduction methods.

In this project, these techniques were applied to a complex chemical descriptor matrix consisting of 150 features and over 10000 samples to achieve optimal dimensionality reduction in preparation for the exploration of any hidden trends and chemical structures within the chemical space.

2.1.1 Data Preparation

Before performing PCA, the dataset was screened to identify any missing values that could potentially skew the results of the analysis. Afterwards, the Docking Scores column, also known as the outcome variable, was extracted from the main dataset to ensure an unbiased representation of the data. Non-numeric data such as the “ID”, “SMILES”, and “target” columns were filtered out of the dataset. Numeric columns with constant data (1 or 0) including the “ComponentCount”, “PosCount”, “NegCount” columns were also removed from the dataset. The retained data (numeric columns with a notable variance structure) and the extracted “Score” column were then separately scaled using the Z-score Normalization method to reduce the impacts of extreme values on the PCA results.

2.1.2 PCA

Using the “prcomp” function in R, PCA was performed on the scaled numeric features to determine the optimal number of principal components. The output of the analysis was validated by reviewing the explained variance and cumulative variance of all the selected principal components.

2.1.3 Elbow Curve

The elbow method is widely employed to predetermine the number of optimal clusters for performing k-means clustering. In this project, the elbow method was used to calculate the Within-Cluster Sum of Squares (WCSS) measure which evaluates the distribution of data points within each cluster. Different k values ranging from 1 to 10 were also selected and iterated over to calculate the WCSS. Following the computation, a graph was plotted to identify the bending point which indicates the optimal number of clusters for the clustering algorithm.

2.1.4 K-means Clustering

K-means clustering is an accompanying unsupervised learning method that is used to separate groups of data with shared similarities into clusters. Once the optimal number of k or clusters were determined from the elbow method, the k-means function was applied to the PCA data. The distance between the data points and the centroids of each cluster were then evaluated. At the end of the process, the data points were assigned to the clusters that had the centroid closest to them.

2.1.5 Uniform Manifold Approximation and Project (UMAP)

UMAP was performed on a denoised dataset, consisting of 25 principal components to enhance computational efficiency. The selected number of principal components was determined based on the total number of principal components with high eigenvalues that accounted for at least 85% cumulative variance of the data's structure. After the importation of the "uwot" package in R, the "n_neighbors" and "min_dist" parameters were then configured to optimally capture global coherence and cluster distribution.

2.2. Regression Analyses

Regression analyses help us predict continuous values. They are a way to estimate the relationship between an outcome variable (dependent variable) and one or multiple predictors (independent variable/s). The most common form of regression analysis is linear regression, where a linear relation between each variable is assumed. It attempts to identify the association between variables by fitting the data into a linear equation.

While it is relatively fast and easily understandable, it risks overfitting and can be sensitive to outliers. Another method of regression is random forest regression, which employs decision trees to draw conclusions from data. Although it is computationally costly, it is a more robust method to determine the relationships between variables and able to capture non-linear relationships.

In this project, both methods of regression were comparatively applied. Before applying regression analyses, the dataset was downsized to 10000 columns for efficiency.

2.2.1. Linear Regression

A simple linear regression model was applied on the entire dataset by first identifying the columns with chemical features. Columns with values of only 0 or 1 were eliminated from the dataset due to their lack of variability decreasing their predictive power. The “caret” library was utilized for the model training: Using the “createDataPartition” function, the modified dataset was separated into training and test datasets following the 80-20 rule. The model was then trained with 5-times cross-validation and with the “lm” method using the “train” function. After summarizing and calculating the RMSE for the test data, the most important 10 chemical features were identified by running random feature selection on the linear regression model. The same procedure described above was applied to fine tune the results.

2.2.2. Random Forest Regression

In addition to linear regression, a random forest regression model was also applied on the dataset. The “rf” method was applied to train the model, with the “mtry” parameter set to a constant value of 4 instead of “caret” to automatically evaluate multiple values, thereby increasing computational cost. Similarly, the number of decision trees to be generated was set to 100 to accelerate model training.

2.2.3. Feature Selection and Error Calculation on Both Regression Models

The “varImp” method was applied to detect the most important 10 chemical features in both models. Root mean squared error (RMSE) and mean absolute error (MAE) values were also calculated for comparative reasons. For time and efficiency purposes, the random forest model was also trained with the random feature elimination control used for linear regression. A density plot was also generated to assess the data distribution after feature selection.

3. Results and Discussion

3.1. Principal Component Analysis (PCA) and K-means Clustering Results

3.1.1.Docking scores across the Chemical Space using PCA

This analysis explored the range of the docking scores in the chemical space as represented by the first two principal components. Docking scores greater than 0, in purple, indicate molecular compounds with a weaker tendency to bind to the enzyme, adenosine deaminase, while the docking scores below -1, represented in yellow, signify molecular compounds with a stronger binding potential. The dense and amorphous shape of the plot can be attributed to the loss of information from the chosen number of principal components, which only encompassed 25% of the data's variance. This suggests that PCA may not be the best dimensionality reduction technique in this use case as it fails to represent the chemical diversity and complex structures present within the data.

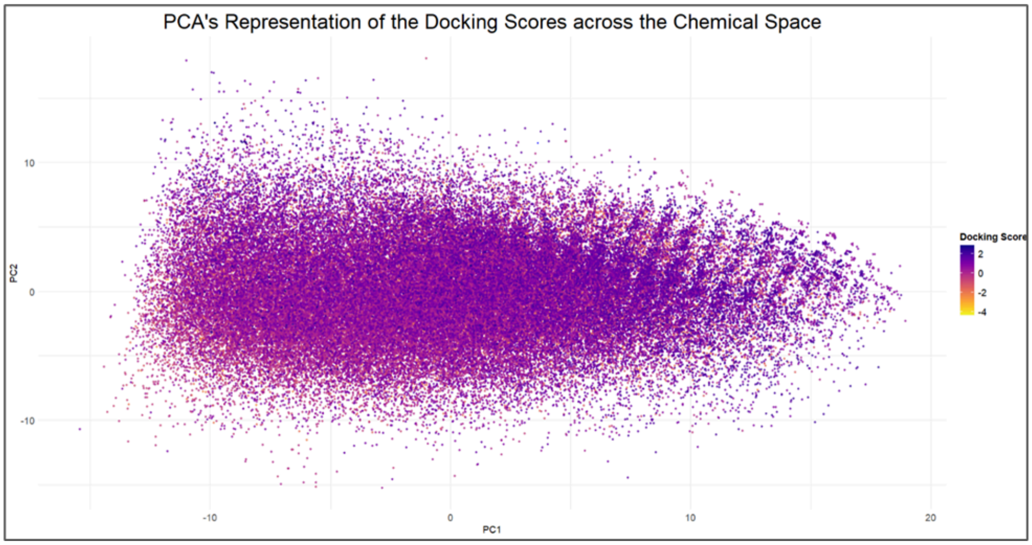

Figure 1. PCA plot illustrating the distribution of the docking scores across the reduced chemical space

3.1.2. Docking scores across the Chemical Space using UMAP

Comparatively, UMAP's visualization of the docking scores depicts a more interpretable global representation of the data's structure as evidenced by its distinct clusters. Small molecules characterized by weaker binding capacity, ranging from 0 and above, are noticeably represented in orange and red on the color scale within each cluster. Conversely, small molecules with stronger binding capacity, ranging from -2 and below, are represented in green and blue colors. The result of the UMAP graph details a more holistic representation of the distribution of the molecules and their apparent underlying relationships due to its segregation of dissimilar molecules and clustering of molecules that may share similar physicochemical properties. Data points that are far isolated from the main clusters indicate anomalies or outliers that require closer examination**.**

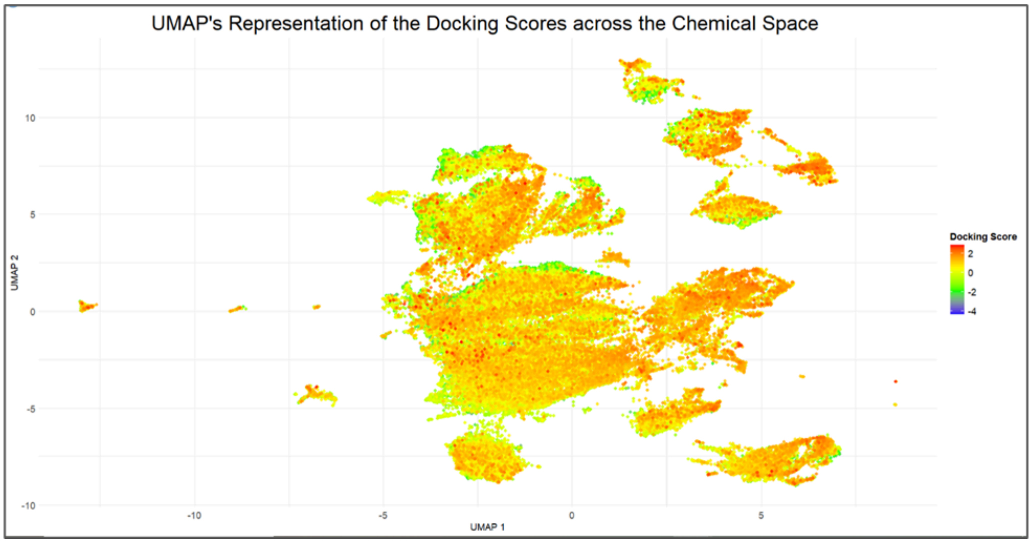

Figure 2. UMAP’s visualisation of the docking scores spread across the chemical space

3.1.3. Clusters across the Chemical Space using PCA

The analysis of the distribution of clusters across the chemical space revealed ten color-coded clusters as shown in the figure below. However, due to the large size of the feature space, the k-means generated clusters largely overlap, highlighting the underlying limitations of PCA as a representative method for the clusters in a low dimensional graph. This finding demonstrates that the molecules in the chemical space may be difficult to group based on shared features due to little to no separability between the measured variables.

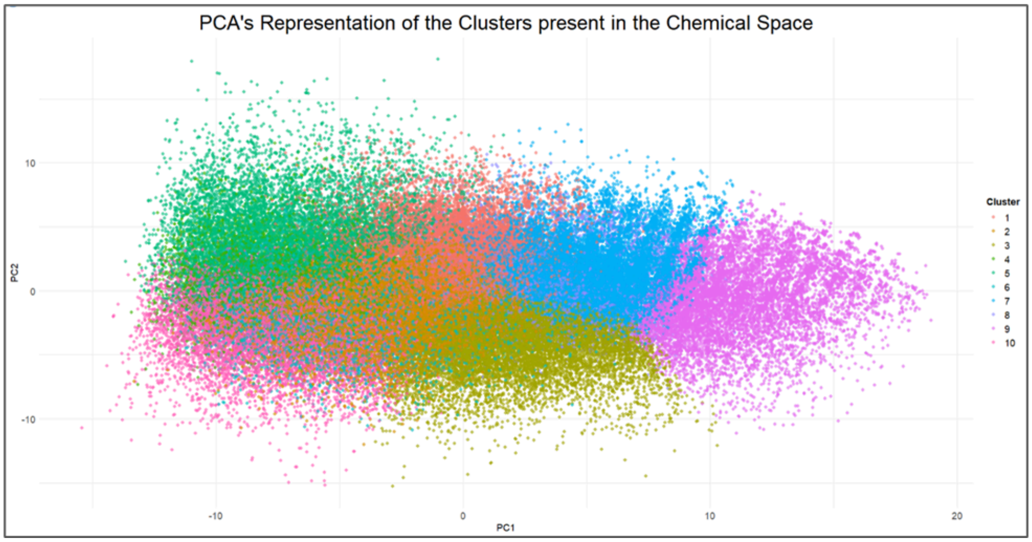

Figure 3. PCA’s distribution of the clusters within the chemical space

3.1.4 Clusters across the Chemical Space using UMAP

In this analysis, traditional clustering with k-means was performed on the PCA-reduced data and visually projected using UMAP to maintain geometric meaning. As shown in the figure below, a close inspection of the plot reveals the fragmentation of identical sub-clusters across different independent clusters, obscuring any cohesive relationship that can be drawn from the observation of both cluster types. This may be as a result of UMAP’s method of preserving local relationships which emphasizes the connectivity of each datapoint to its nearest neighbour. As such, data points with differing neighbor connectivity are more likely to appear fragmented across the UMAP plot. Alternatively, k-means’ assumption that clusters within the same Euclidean space are spherical, and isotropic may have also impacted the repetition or partitioning of subclusters.

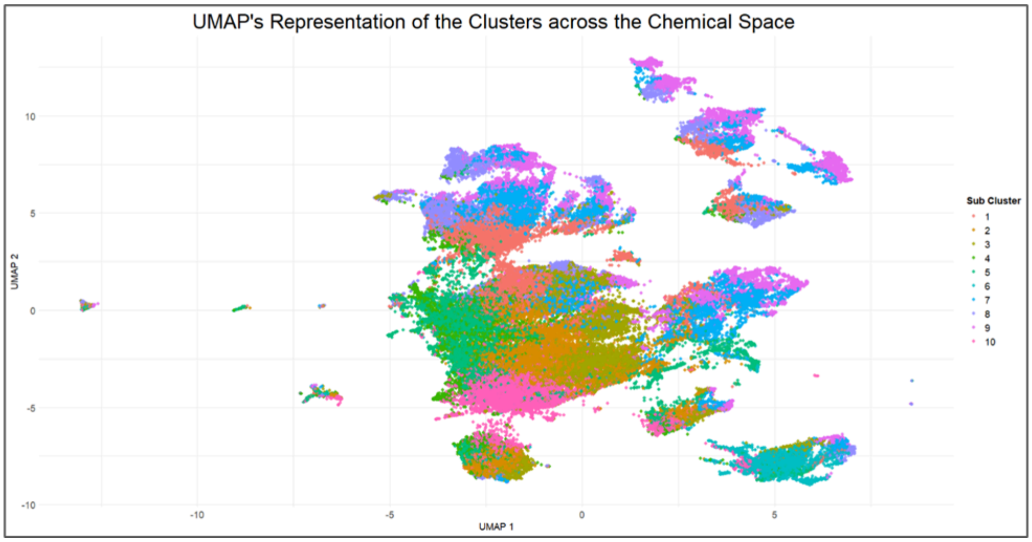

Figure 4. UMAP's visualisation of the k-means generated clusters across the chemical space

3.2. Regression Analyses

3.2.1. Linear Regression

Linear regression analysis on the modified dataset revealed a low RMSE of ~22.03. Based on the linear regression model generated, chemical features alone can be used to predict the docking score to some extent, despite the R^2 value being low. However, this points to possible noise in the data, or that another regression model needs to be applied (e.g.: Random forest regression). Interestingly, each time the model was run, the most important features varied greatly. Choice of regression model depended on the computing capacity and time cost. The 10 most important chemical features detected are given in Figure 4.

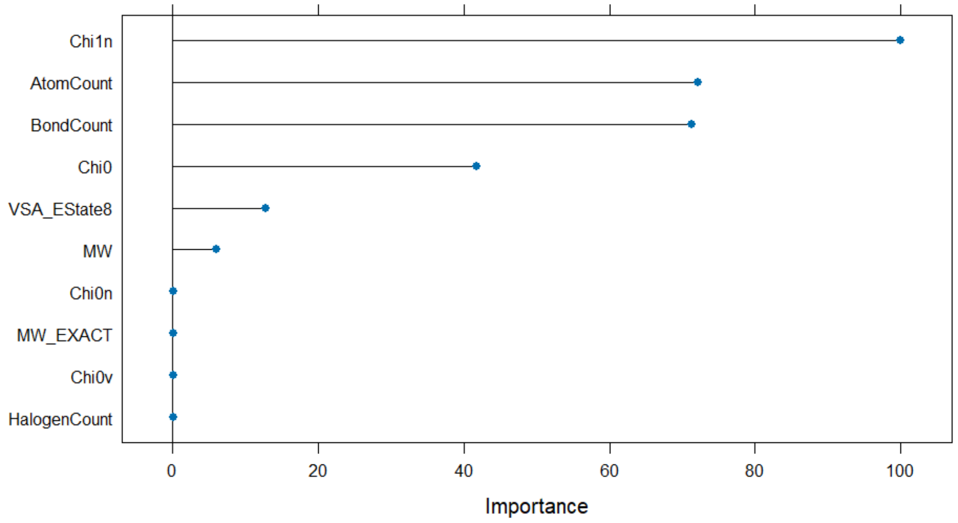

Figure 5. The final plot of the 10 most important chemical features for the detection of docking scores, based on the linear regression model.

As seen in Figure 4, the most important 10 chemical features include the topological/topochemical features “Chi1n”,”Chi0” and “Chi0v”, physical features like the “AtomCount”, “BondCount”, “MW”, “MW_EXACT” and “HalogenCount”, as well as the electrotopological state descriptor “VSA_EState8”.

3.2.2. Random Forest Regression

Random forest regression identified the RMSE as ~25.52, higher than linear regression. 4 specific features were identified as the most important chemical features contributing to the docking scores: “Chi0n”, “Chi0v”, “HeavyAtomMolWt” and “MW_EXACT”.

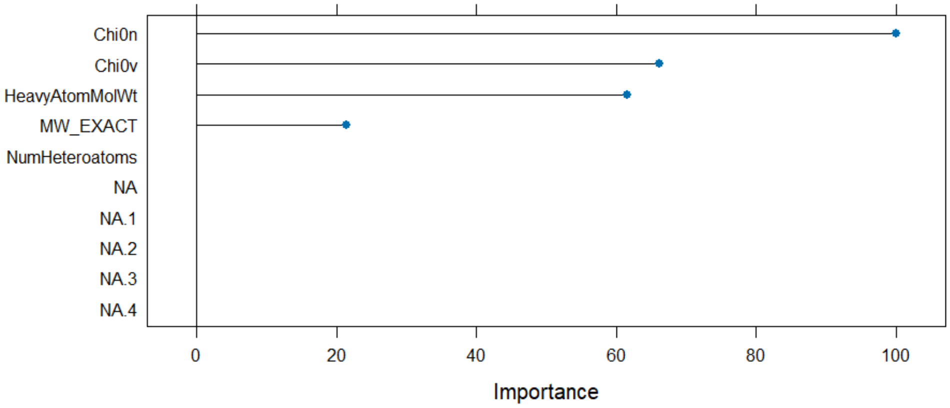

Figure 6. The final plot of the 10 most important chemical features for the detection of docking scores, based on the random forest regression model.

MAE, RMSE, R^2 values and the data distributions before and after feature selection were plotted as well:

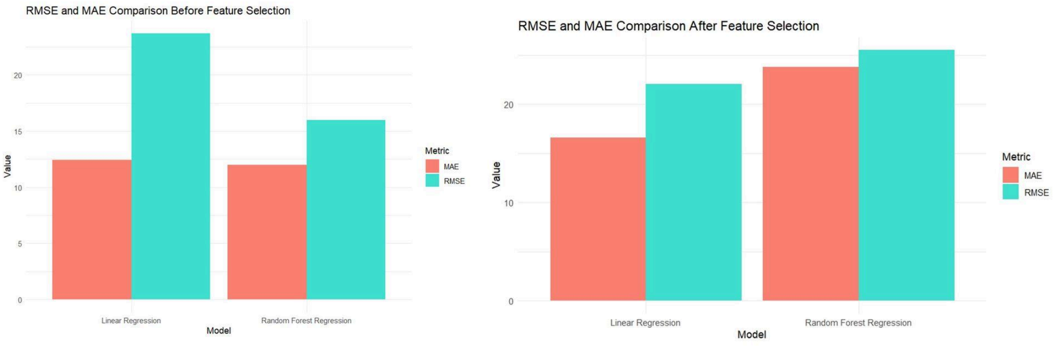

Figure 7. RMSE and MAE values before and after feature selection for both models.

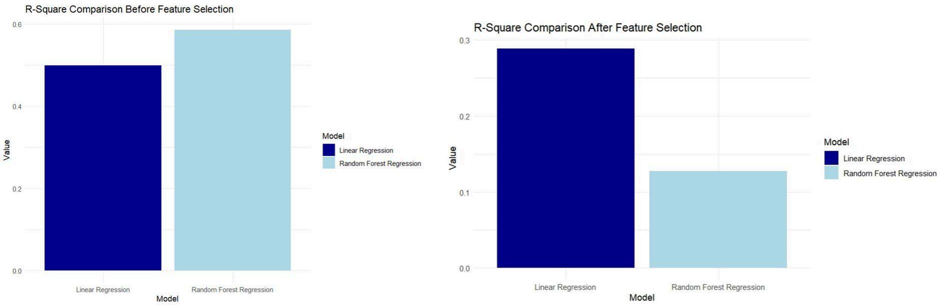

Figure 6. R^2 values before and after feature selection for both models.

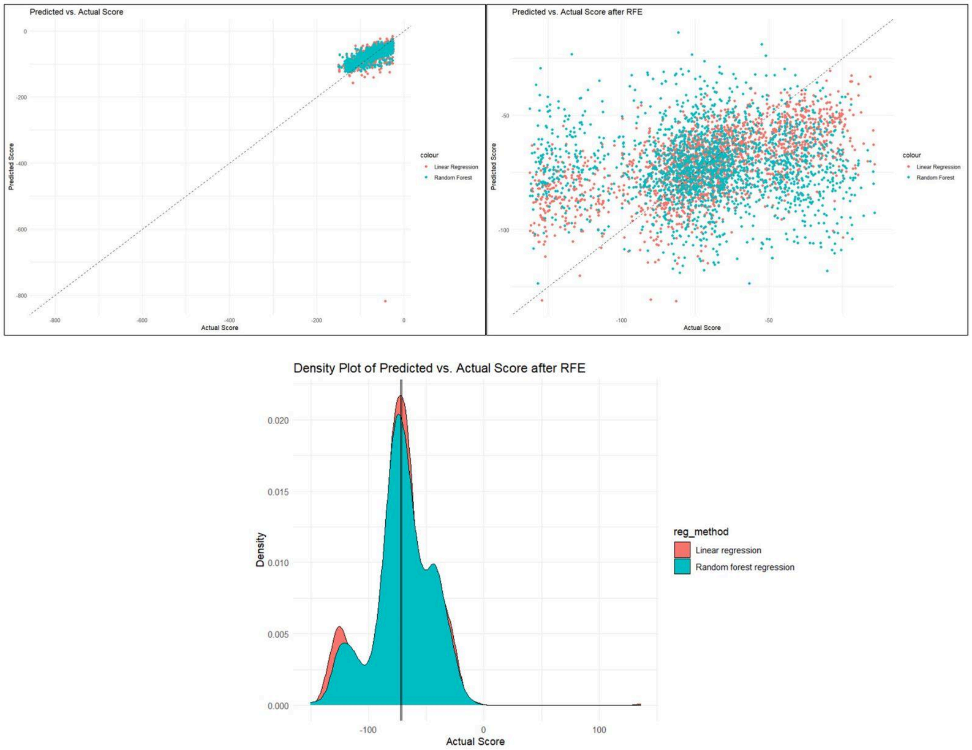

Figure 8. Scatter plots before and after feature selection for both models and density plot of scores after recursive feature selection.

4. Conclusion

Current results from the regression analyses showed increased variations in the most important chemical features in determining the docking scores, possibly due to pre-processing and the random sampling of the data and the variability in the features themselves. The increase in RMSE and MAE after feature selection in random forest regression, as well as the decrease in R^2 values could be a result of removing variables with constant values or the choice of parameter values used in recursive feature elimination (RFE). Additionally, using the same RFE control for both random forest regression and linear regression could have decreased the accuracy of the models.

Based on the visualisation results of the dataset's chemical space, PCA can be regarded as an unsuitable dimensionality reduction technique in this use case as a result of its inability to fully capture all the complex, non-linear relationships available in the data. UMAP's inherent separation ability, however, enables it to render a better structural preservation of the variance present in the high dimensional dataset.

Despite UMAP's ability to retain the global structure of the dataset, its graphical representation is not inherently chemically meaningful due to its unsupervised nature. Likewise, the new, orthogonal features extracted by PCA may not be visibly interpretable. It is, therefore, imperative that the results of both models are further analyzed and validated in order to ascertain more relevant information from the embeddings of the plots.

Exploring the factor loadings from the PCA results may uncover the correlation between each variable and the principal components. The Identification of the structural classes of the compounds within each cluster represented in the UMAP's plot can also reveal useful insights about the reliability of global distances and its implications for the positioning of each cluster. Furthermore, it can aid the interpretation of the extent of local diversity represented by the spread of a cluster and the relationship between clusters based on their proximity.

5. References

- Zavialov, A.V., Gracia, E., Glaichenhaus, N., Franco, R., Zavialov, A.V. and Lauvau, G. 2010. “Human adenosine deaminase 2 induces differentiation of monocytes into macrophages and stimulates proliferation of T helper cells and macrophages”. Journal of Leukocyte Biology, 88(2), 279-290.

- Grisoni, F., Reker, D., Schneider, P., Friedrich, L., Consonni, V., Todeschini, R., Koeberle, A., Werz, O. and Schneider, G. 2016. Matrix-based Molecular Descriptors for Prospective Virtual Compound Screening. Molecular Informatics, 36(1-2). PMID: 27650559.