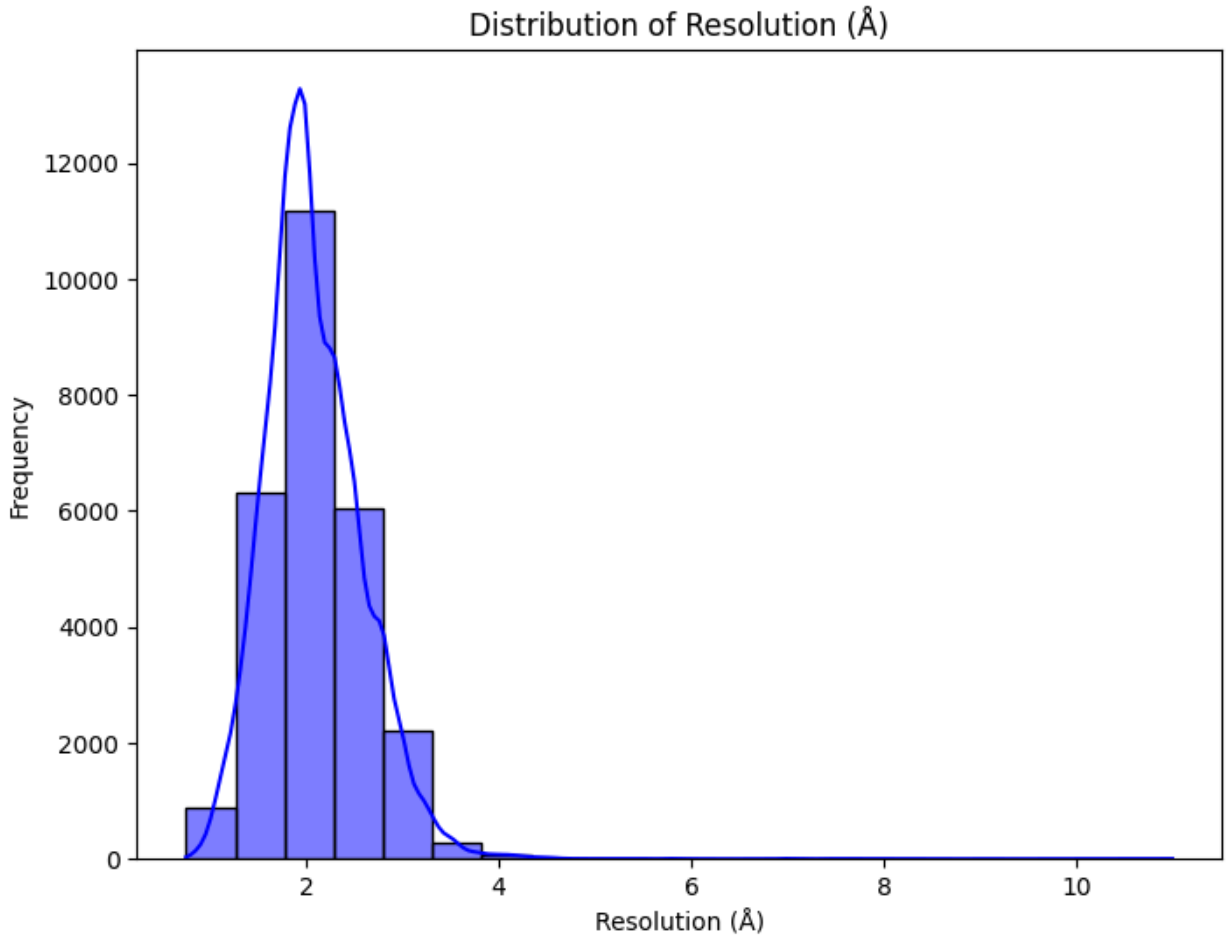
Graphs with explanation:

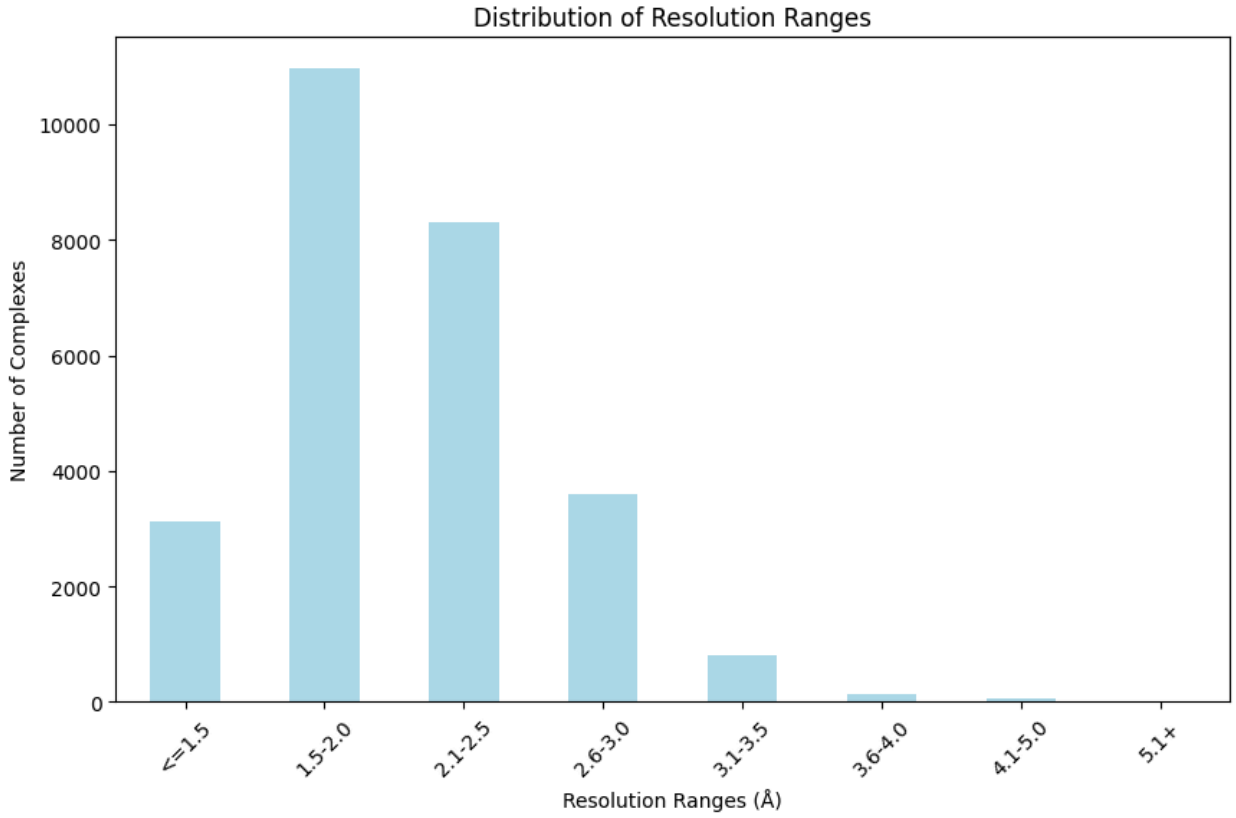


The blue bars represent the histogram. The height of each bar corresponds to the frequency of data points within the corresponding resolution range (bin). The histogram reveals the general shape of the data distribution. In this case, it appears to be skewed to the right (positively skewed). This means there's a long tail on the right side of the distribution. The highest bar indicates the most frequent resolution range. The peak is around 2 Å, suggesting that a large number of data points have a resolution close to this value. The bars show the range of resolutions present in the data. We can see that most resolutions fall between 1 Å and 4 Å, but there are some values extending beyond 10 Å.

The blue curve superimposed on the histogram is likely a kernel density estimation (KDE). This curve provides a smoothed representation of the data distribution. The curve reinforces the observation of the right-skewed distribution. It also highlights the peak around 2 Å more clearly. The curve helps visualize the overall trend of the data, smoothing out the "jaggedness" of the histogram bars.

The peak of the distribution suggests that the most common resolution in the dataset is around 2 Å. The concentration of data points in the lower resolution range (1-4 Å) indicates that the dataset is dominated by relatively high-resolution values. The long tail extending to higher resolution values (beyond 4 Å) indicates the presence of some data points with lower resolution.

This graph depicts the distribution of resolution values in a dataset. The data exhibits a right-skewed distribution, with a peak around 2 Å, indicating that the dataset consists primarily of high-resolution data points. However, there is a presence of lower-resolution data, as evidenced by the tail extending to higher resolution values. The graph provides a clear visual representation of the typical resolution and the spread of resolution values in the dataset. This graph tells us that most of the data is very detailed (high resolution), but there are some less detailed pieces of data mixed in.

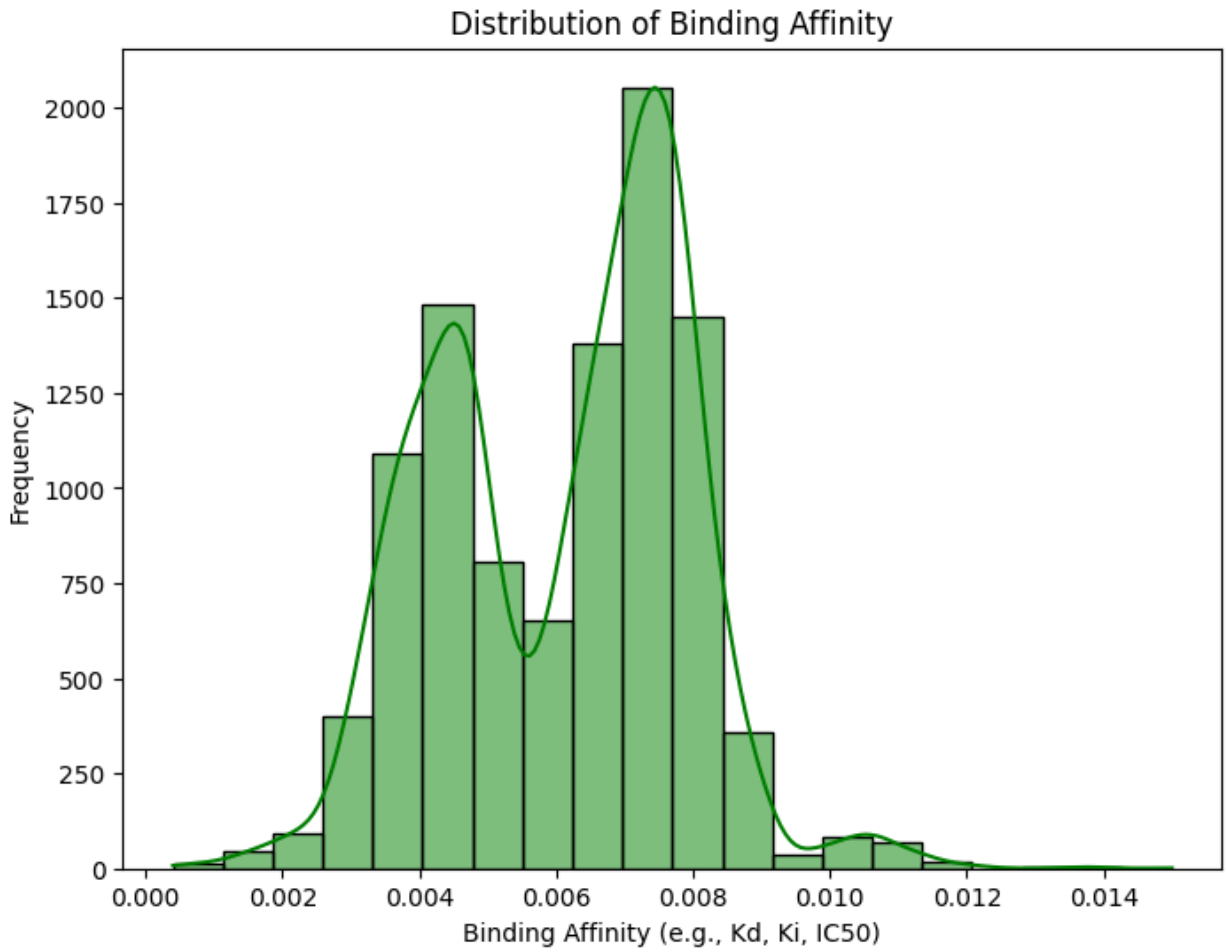


The distribution shows a clear trend. The number of complexes is high for lower resolution values (higher detail) and decreases as resolution values increase (lower detail). This suggests a preference or bias towards higher-resolution data in the dataset. A steep decrease in the number of complexes as we move from the 1.5-2.0 Å range to the 2.1-2.5 Å range, and a continued decrease in subsequent ranges. The bars for higher resolution ranges (3.1-3.5 Å and above) are very short, indicating a very low number of complexes with lower resolution.

The graph shows a clear dominance of high-resolution data in the dataset, with the majority of complexes having resolutions between 1.5 and 2.5 Å. The distribution suggests a potential bias

towards collecting or publishing data with higher resolution, or a limitation in obtaining high-quality data at lower resolutions. The most common resolution range for the complexes in this dataset is 1.5-2.0 Å. There are very few complexes with resolutions above 3.0 Å, indicating that lower-resolution data is less common or less represented in the dataset.

This graph presents the distribution of resolution ranges for a set of complexes. The data shows a strong bias towards high-resolution structures, with the majority of complexes having resolutions between 1.5 and 2.5 Å. The number of complexes decreases significantly as resolution decreases, suggesting a preference or limitation in obtaining high-quality data at lower resolutions. The graph effectively highlights the trend of structural biology research towards achieving and utilizing higher-resolution data. Most of the molecular structures in the dataset are very detailed (high resolution), and there are very few structures with lower detail (lower resolution).

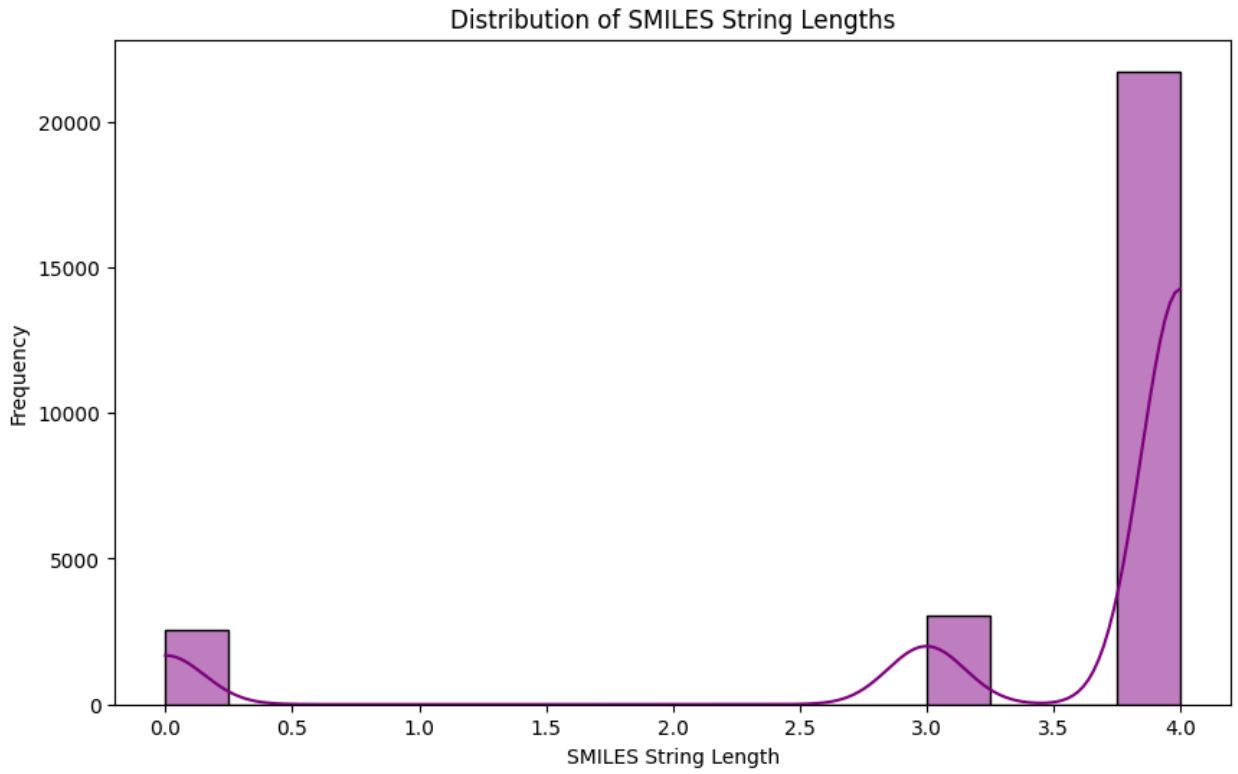


The green curve superimposed on the histogram is likely a kernel density estimation (KDE). This curve provides a smoothed representation of the data distribution. The curve reinforces the

observation of the bimodal distribution. It clearly highlights the two peaks and the valley between them. The curve helps visualize the overall trend of the data, smoothing out the "jaggedness" of the histogram bars.

The two peaks indicate a significant difference in binding affinity between the two groups. One group shows higher binding affinity (lower values), suggesting a stronger interaction, while the other group shows lower binding affinity (higher values), suggesting a weaker interaction. The bimodal distribution might have biological significance. It could indicate different functional states of a protein, different binding partners, or different types of ligands.

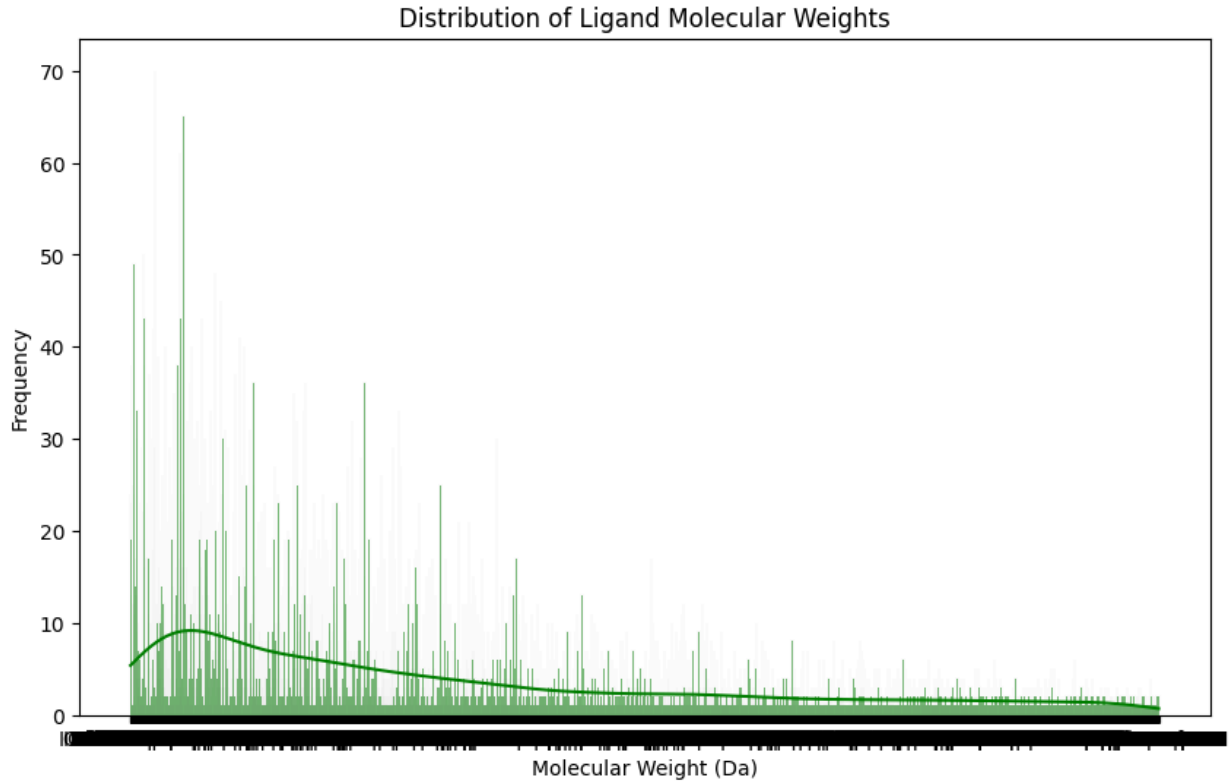
This graph depicts the distribution of binding affinity values, revealing a bimodal distribution. The presence of two distinct peaks suggests that the data likely comes from two different populations or reflects two different binding modes. This indicates a significant difference in binding affinity between the two groups, with one group showing higher binding affinity and the other showing lower binding affinity. The bimodal distribution might have biological significance and could indicate different functional states, binding partners, or types of ligands. There are two main groups of interactions with different strengths. Some interactions are strong (high binding affinity), and others are weaker (low binding affinity).



The dark purple curve superimposed on the histogram is likely a kernel density estimation (KDE). This curve provides a smoothed representation of the data distribution. The trimodal distribution suggests that the data likely comes from three distinct groups of molecules based on their complexity or size. The largest peak at 4 indicates that the majority of SMILES strings in

the dataset are of length 4. This suggests a potential bias or selection towards molecules with a specific level of complexity. The peaks at 0 and 3 indicate the presence of molecules with simpler structures (short strings) and molecules with intermediate complexity (medium length strings).

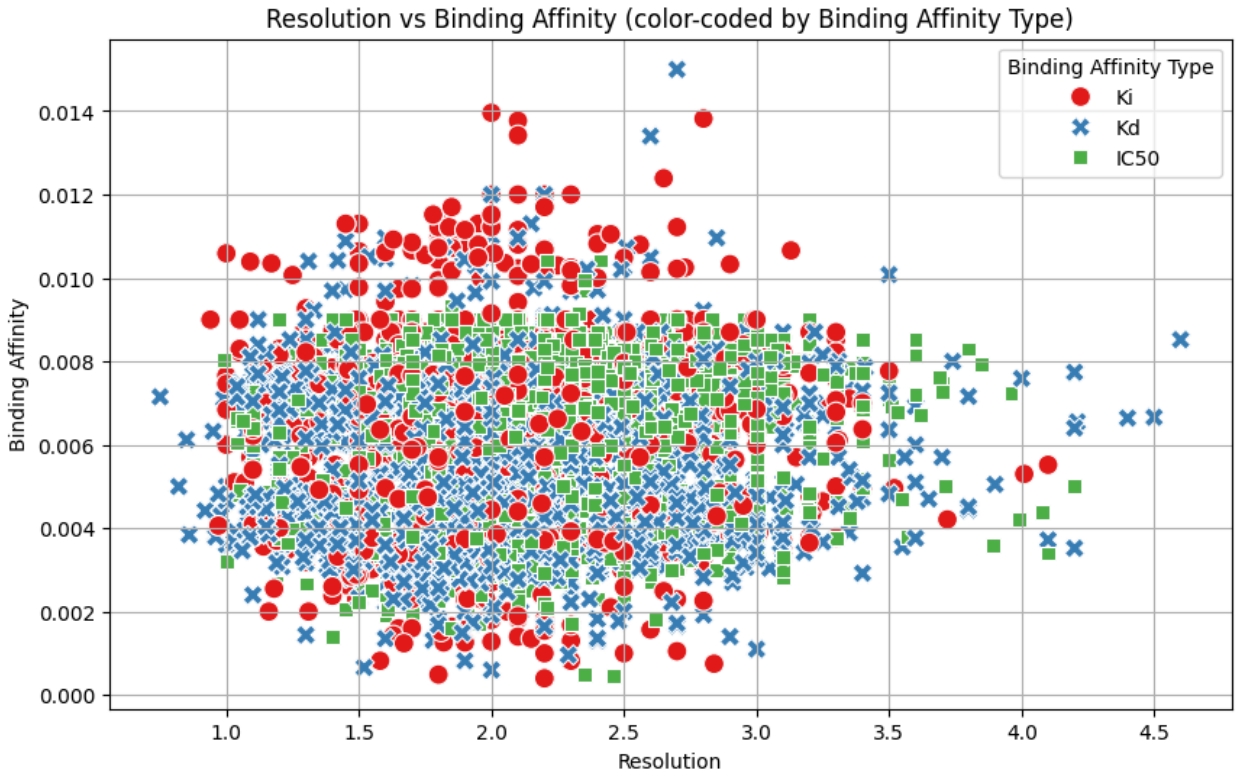
This graph depicts the distribution of SMILES string lengths, revealing a trimodal distribution. The presence of three distinct peaks suggests that the data likely comes from three different groups of molecules based on their complexity. The majority of the molecules have SMILES strings of length 4, indicating a potential bias or selection towards molecules with a specific level of complexity. The graph also shows the presence of molecules with simpler structures (short strings) and molecules with intermediate complexity (medium length strings). The dataset contains molecules of three main levels of complexity, with the majority being of a specific complexity (length 4).



The graph shows a clear dominance of ligands with low molecular weights in the dataset.

This graph depicts the distribution of ligand molecular weights, revealing a heavily right-skewed distribution. The data shows a clear dominance of ligands with low molecular weights, with the majority concentrated at the lower end of the molecular weight spectrum. The number of ligands decreases significantly as molecular weight increases, indicating that high molecular weight ligands are less common in the dataset. The graph suggests a potential bias or selection towards low molecular weight ligands, or a limitation in obtaining or studying high molecular

weight ligands. Most of the ligands in the dataset are small molecules (low molecular weight), and there are very few large molecules (high molecular weight).

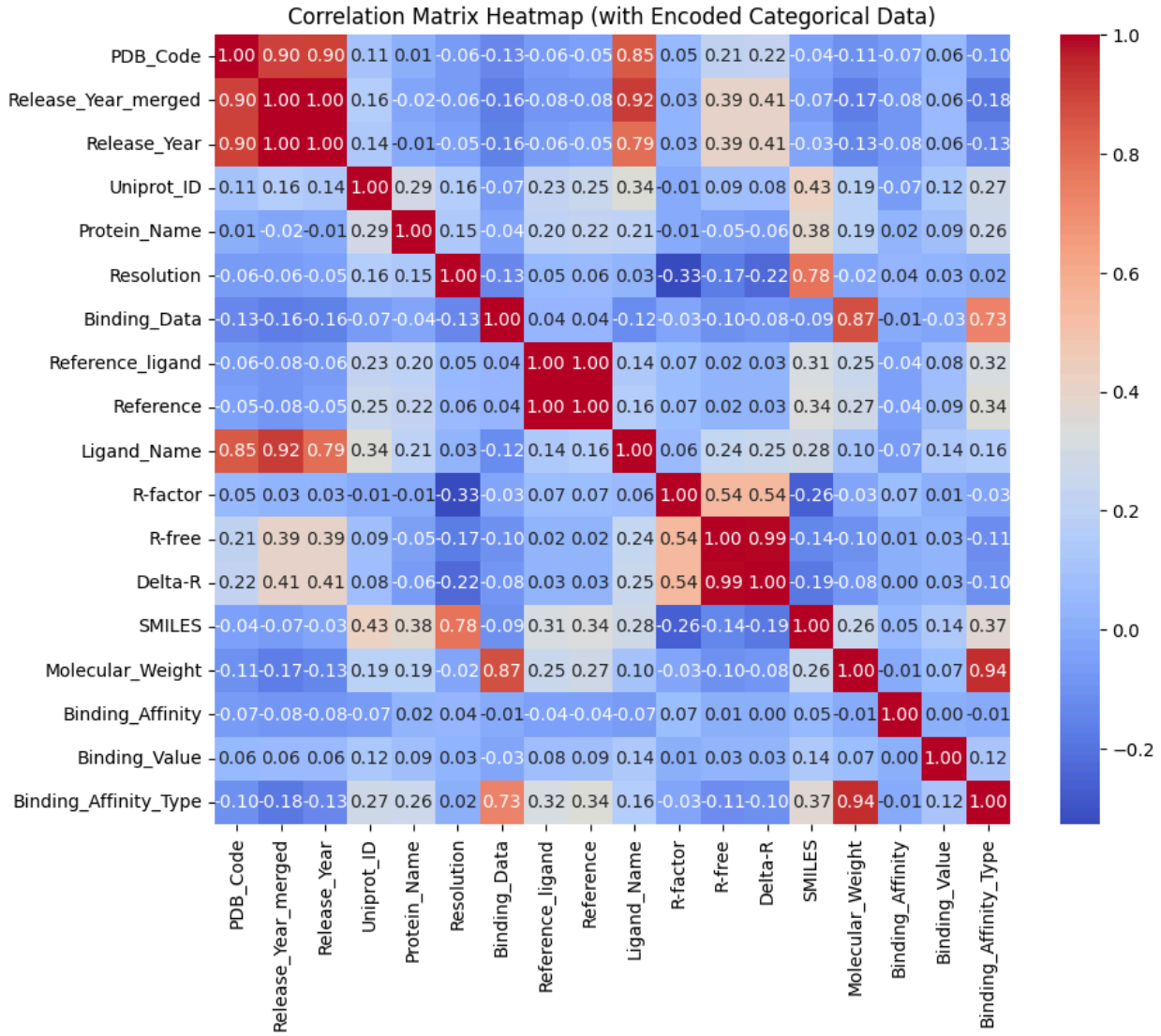


There doesn't appear to be a strong, obvious correlation or trend between resolution and binding affinity across the entire dataset. The points are relatively evenly distributed across the range of resolutions, with no clear pattern of binding affinity changing with resolution. The points representing different binding affinity types (Ki, Kd, IC50) are interspersed and overlap significantly. This suggests that there isn't a clear separation or clustering of data points based on the type of binding affinity measurement. The distributions of Ki, Kd, and IC50 values across the range of resolutions appear to be similar. This indicates that the choice of binding affinity measurement might not strongly influence the observed relationship (or lack thereof) between resolution and binding affinity.

The lack of a clear trend suggests that there is no strong direct relationship between resolution and binding affinity in this dataset. Higher resolution does not necessarily imply higher or lower binding affinity. Resolution and binding affinity might be independent variables, or their relationship might be more complex and influenced by other factors not shown in the graph.

This scatter plot shows the relationship between resolution and binding affinity for a dataset of molecular interactions, color-coded by the type of binding affinity measurement (Ki, Kd, IC50). The lack of a clear trend suggests that there is no strong direct correlation between resolution and binding affinity in this dataset. The data points are scattered across the graph, indicating significant variability in both resolution and binding affinity. The different binding affinity types are

interspersed, suggesting that the choice of measurement might not strongly influence the observed relationship. There's no clear pattern between how detailed the data is (resolution) and how strongly molecules interact (binding affinity). The strength of the interaction doesn't seem to depend on how detailed the data is.



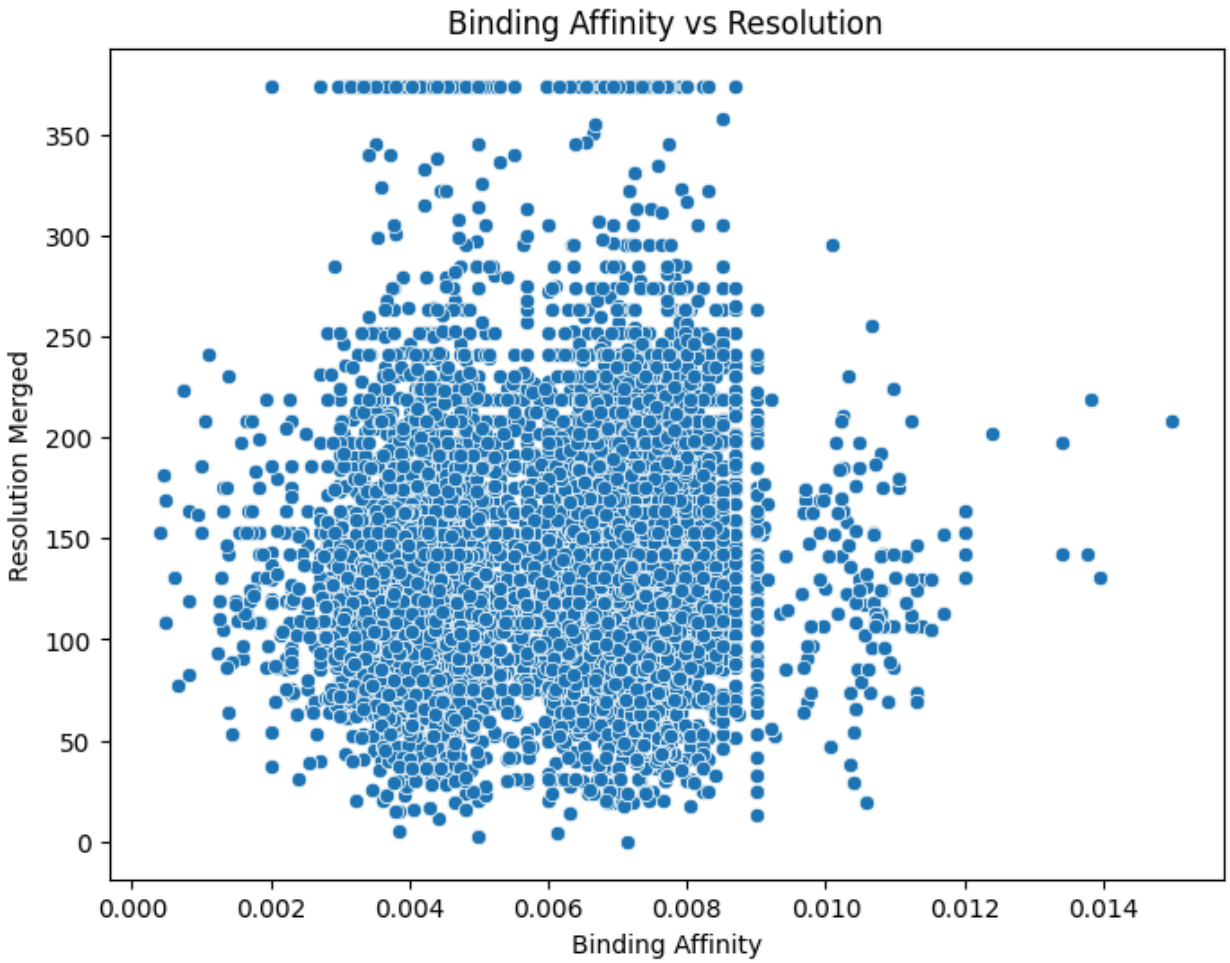
Strong Positive Correlations: `Release_Year_merged` and `Release_Year` (1.00): This indicates that these two variables are perfectly correlated, likely meaning they represent the same information or are derived from each other. `R-free` and `Delta-R` (0.99): This indicates a very strong positive correlation, suggesting that as `R-free` increases, `Delta-R` also tends to increase. `Molecular_Weight` and `Binding_Affinity_Type` (0.94): This indicates a strong positive correlation, suggesting that higher molecular weights are associated with a particular type of binding affinity measurement. `PDB_Code` and `Ligand_Name` (-0.85): This indicates a

strong negative correlation, suggesting that as PDB_Code increases, Ligand_Name tends to decrease, or vice-versa. (Note: The interpretation of negative correlations with encoded categorical variables can be tricky.) Binding_Data and Molecular_Weight (0.87): This indicates a strong positive correlation, suggesting that higher binding data is associated with higher molecular weights.

Strong Negative Correlations: PDB_Code, Release_Year_merged, and Release_Year (-1.00, -0.90, -0.90): This indicates strong negative correlations between PDB_Code and the release year variables, suggesting that as PDB_Code increases, the release year tends to decrease (or vice-versa).

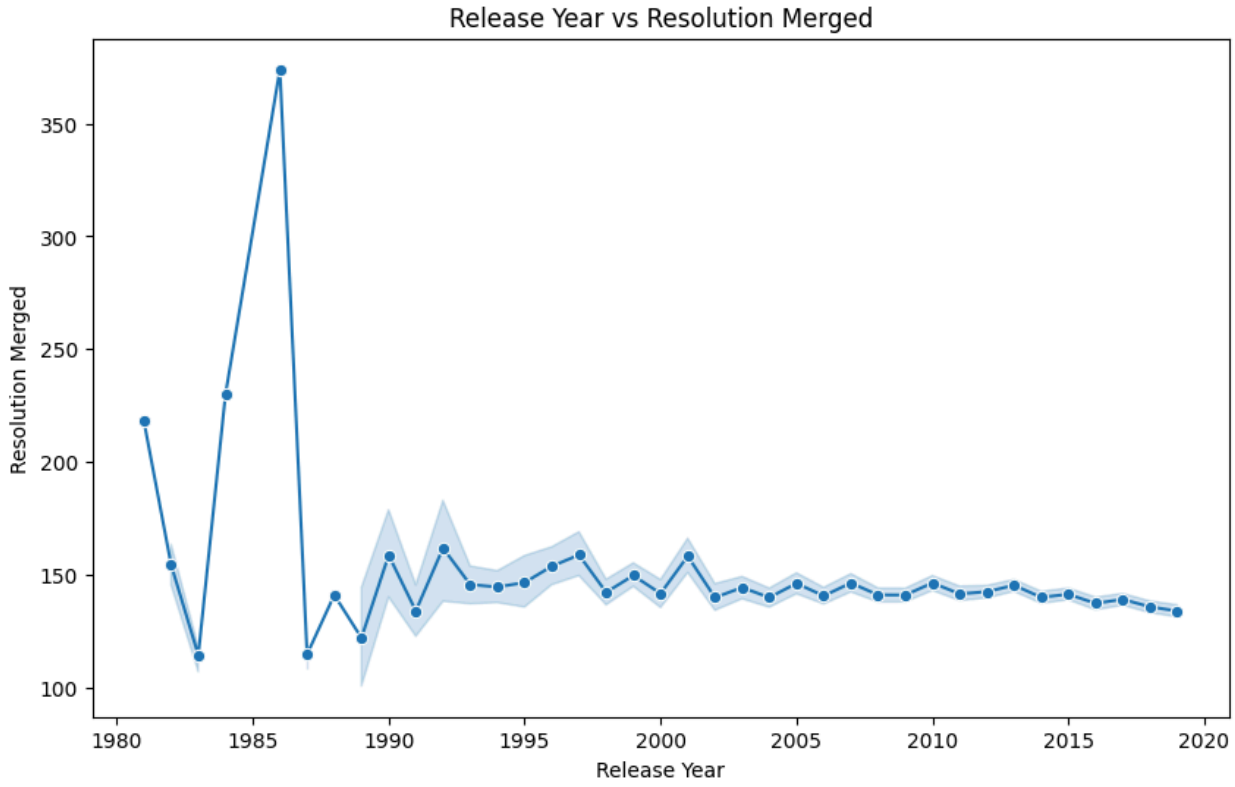
Weak Correlations: Many cells have correlations close to 0 (white or light colors), indicating weak or no linear correlation between those pairs of variables.

The perfect correlation between Release_Year_merged and Release_Year suggests that one of these variables might be redundant. The strong correlations between Molecular_Weight, Binding_Affinity_Type, and Binding_Data suggest potential relationships between these variables that warrant further investigation. The correlation between Resolution and Binding_Affinity is relatively weak (0.04), suggesting that there isn't a strong linear relationship between these variables in this dataset.



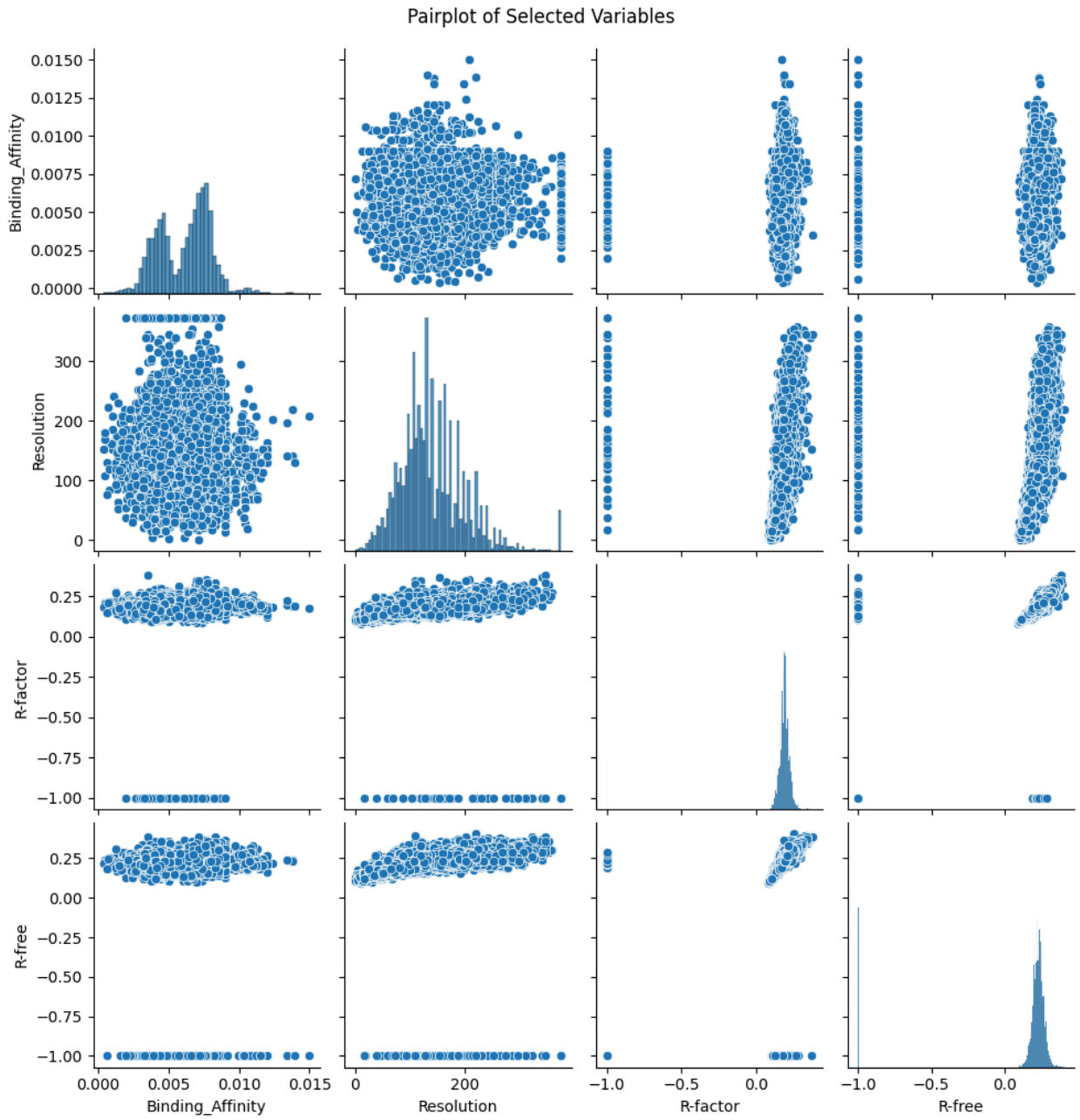
The lack of a clear trend suggests that there is no strong direct relationship between binding affinity and resolution in this dataset. Higher binding affinity does not necessarily imply higher or lower resolution. Binding affinity and resolution might be independent variables, or their relationship might be more complex and influenced by other factors not shown in the graph. The scatter plot indicates significant variability in both binding affinity and resolution. This suggests that other factors, such as experimental conditions, protein structure, or ligand properties, might play a significant role in determining binding affinity and resolution.

This scatter plot shows the relationship between binding affinity and resolution for a dataset of molecular interactions. The lack of a clear trend suggests that there is no strong direct correlation between binding affinity and resolution in this dataset. The data points are scattered across the graph, indicating significant variability in both binding affinity and resolution. The horizontal banding at the top of the plot suggests a potential artifact or data processing issue. There's no clear pattern between how strongly molecules interact (binding affinity) and the quality of the data (resolution). The strength of the interaction doesn't seem to depend on the quality of the data. However, the horizontal banding suggests that there might be some limitations or specific characteristics in the resolution data that need to be considered.



The graph clearly shows an improvement in resolution over time, likely due to advancements in technology and techniques. The higher "Resolution Merged" values and greater variability in early data suggest limitations in data quality and consistency during that period. The rapid improvement in resolution in the late 1980s likely corresponds to significant technological advancements in the field. The plateauing of resolution improvement in recent years suggests that while progress continues, it may be occurring at a slower rate.

This line plot shows the trend of "Resolution Merged" over time, revealing a clear improvement in resolution as technology and techniques have advanced. The early data points show lower resolution and greater variability, while the recent data points show higher resolution and less variability. The graph also suggests a potential plateauing of resolution improvement in recent years. The quality of the data (resolution) has been getting better over time, likely due to improvements in technology. The data was not as good in the early years, but it has improved significantly since then.



The lack of clear relationships between binding affinity and resolution, R-factor, or R-free suggests that the strength of molecular interactions is not strongly dependent on the quality of the structural data in this dataset. The observed trend between resolution and R-factor/R-free confirms the expected relationship between data quality and model quality in structural biology. The strong positive correlation between R-factor and R-free validates the consistency of these model quality measures. The bimodal distribution of binding affinity suggests that the dataset contains interactions with two distinct binding modes or affinities. This could be due to different types of interactions, different binding sites, or different experimental conditions.

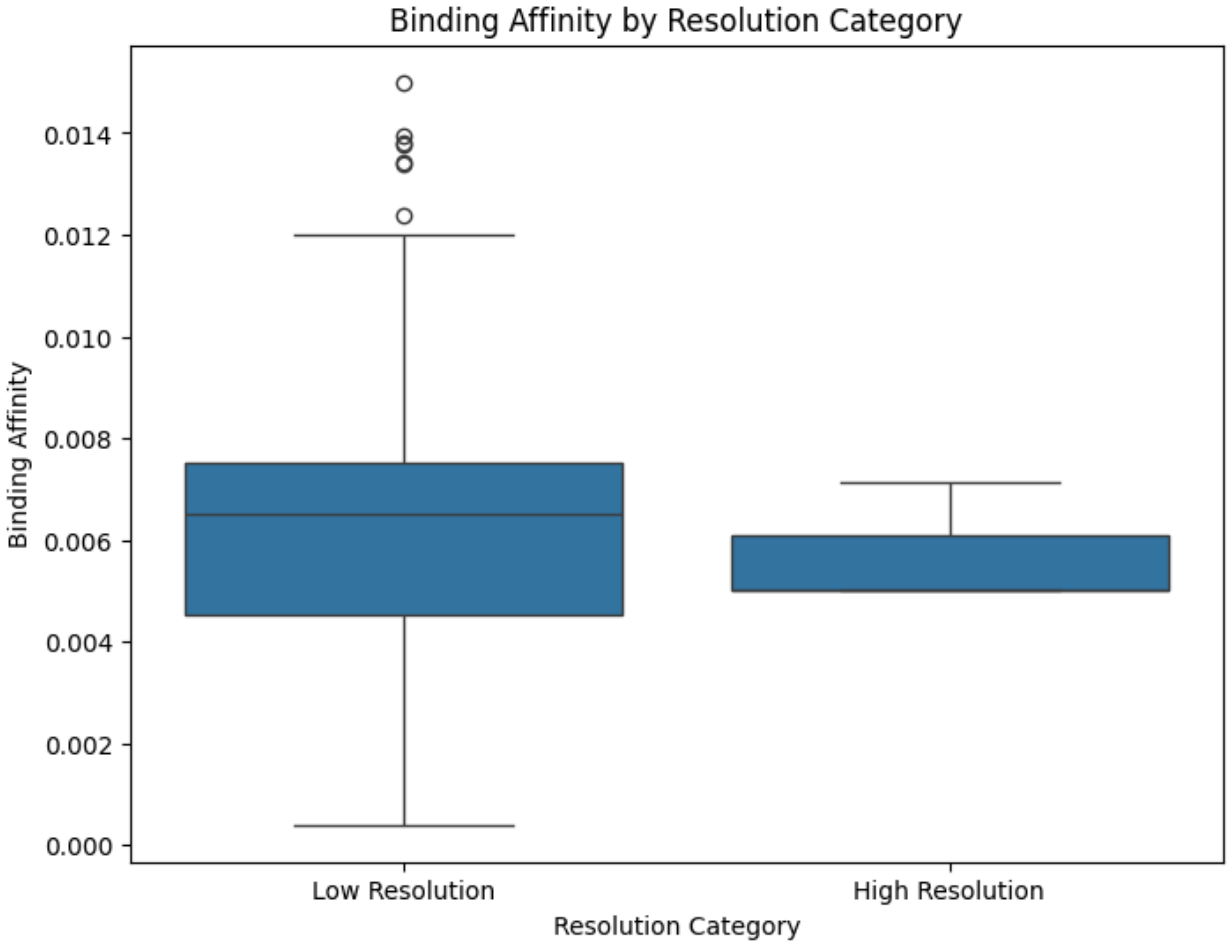
This pairplot provides a comprehensive overview of the relationships between binding affinity, resolution, and model quality measures. It reveals the absence of strong correlations between binding affinity and structural quality, while confirming expected relationships between resolution and model quality measures. The bimodal distribution of binding affinity suggests the presence of distinct interaction modes.

The strength of molecular interactions doesn't seem to depend on the quality of the structural data. Better structural data (higher resolution) generally leads to better models. There are likely two distinct types of molecular interactions in the dataset, based on their binding strength.

The lack of a clear trend suggests that there is no strong direct relationship between resolution and binding affinity in this dataset. Higher resolution does not necessarily imply higher or lower binding affinity. Resolution and binding affinity might be independent variables, or their relationship might be more complex and influenced by other factors not shown in the graph. The scatter plot indicates significant variability in both binding affinity and resolution. This suggests that other factors, such as experimental conditions, protein structure, or ligand properties, might play a significant role in determining binding affinity and resolution.

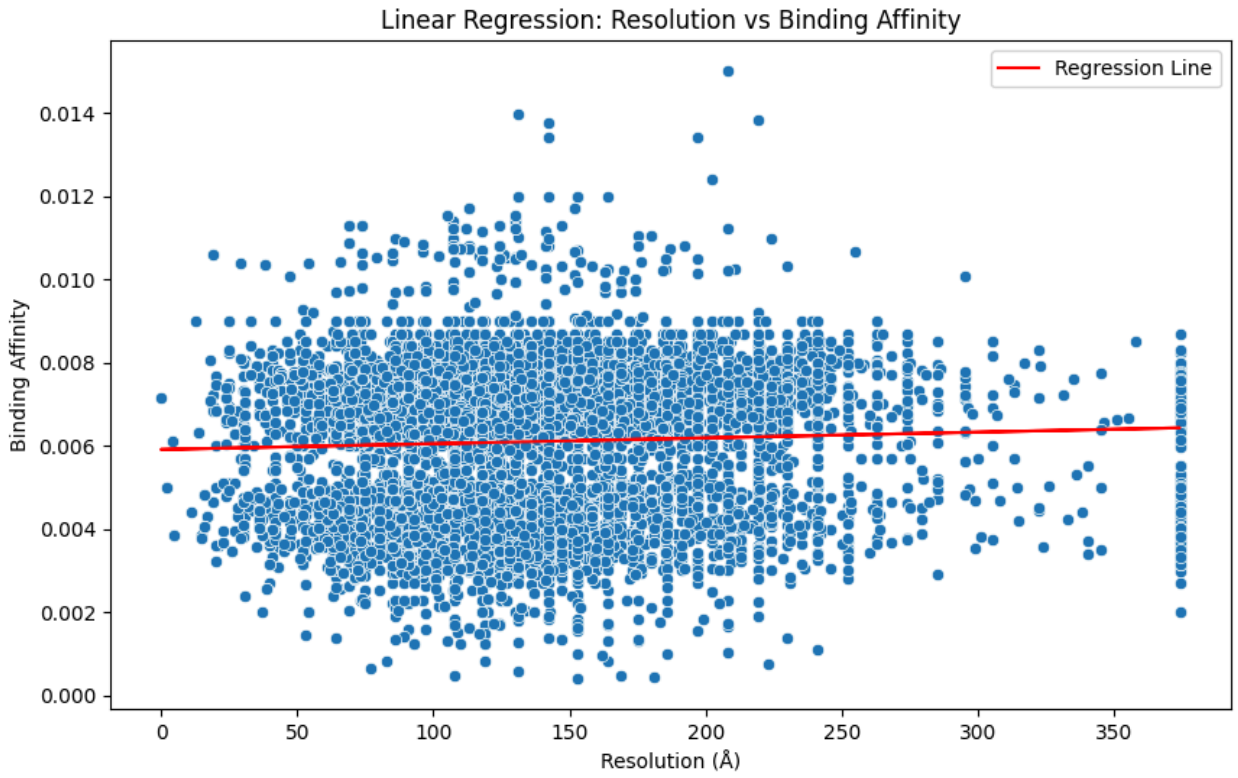
This scatter plot shows the relationship between resolution and binding affinity for a dataset of molecular interactions. The lack of a clear trend suggests that there is no strong direct correlation between resolution and binding affinity in this dataset. The data points are scattered across the graph, indicating significant variability in both binding affinity and resolution. There's no clear pattern between how detailed the data is (resolution) and how strongly molecules

interact (binding affinity). The strength of the interaction doesn't seem to depend on how detailed the data is.



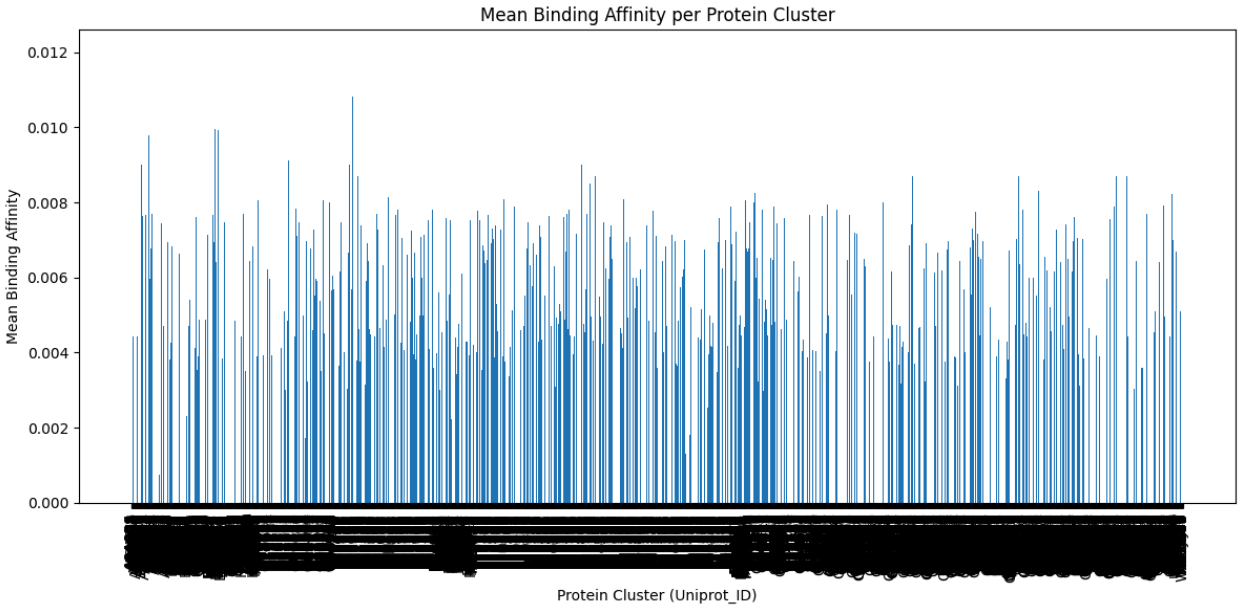
The slightly lower median and narrower IQR for the "High Resolution" category suggest a weak trend where higher resolution data points tend to have slightly higher binding affinities (stronger interactions). The wider IQR and whiskers, along with the presence of outliers, indicate that the binding affinities for the "Low Resolution" data points are more variable and less consistent. The greater variability and outliers in the "Low Resolution" data might suggest potential data quality issues or inconsistencies in this group.

This boxplot compares the distribution of binding affinities between data points categorized as having "Low Resolution" and "High Resolution". The graph reveals a slight trend where higher resolution data points tend to have slightly higher binding affinities (stronger interactions). However, the binding affinities for the "Low Resolution" data points are more variable and less consistent, potentially indicating data quality issues or inconsistencies. While there's a slight tendency for data with better quality (high resolution) to show stronger interactions, the data with lower quality shows a wider range of interaction strengths and potentially some unreliable measurements.



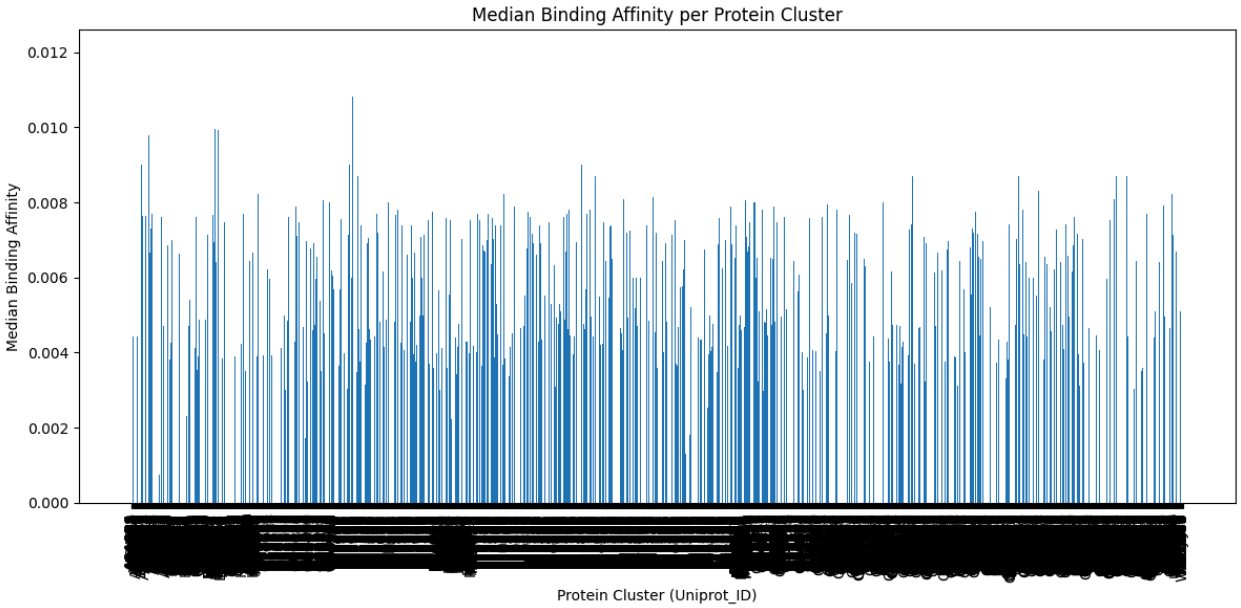
The slightly positive slope of the regression line suggests a very weak positive correlation between resolution and binding affinity. However, the correlation is so weak that it is likely not practically significant. The poor fit of the regression line indicates that a linear model is not appropriate for describing the relationship between resolution and binding affinity in this dataset.

This scatter plot shows the relationship between resolution and binding affinity, with a linear regression line fitted to the data. The regression line suggests a very weak positive correlation, but the poor fit of the line indicates that a linear model is not appropriate for describing the relationship. The wide scatter of data points suggests that other factors play a significant role in determining binding affinity. There's no meaningful relationship between how detailed the data is (resolution) and how strongly molecules interact (binding affinity). The strength of the interaction doesn't seem to depend on how detailed the data is, and a simple straight line is not a good way to describe any potential relationship.

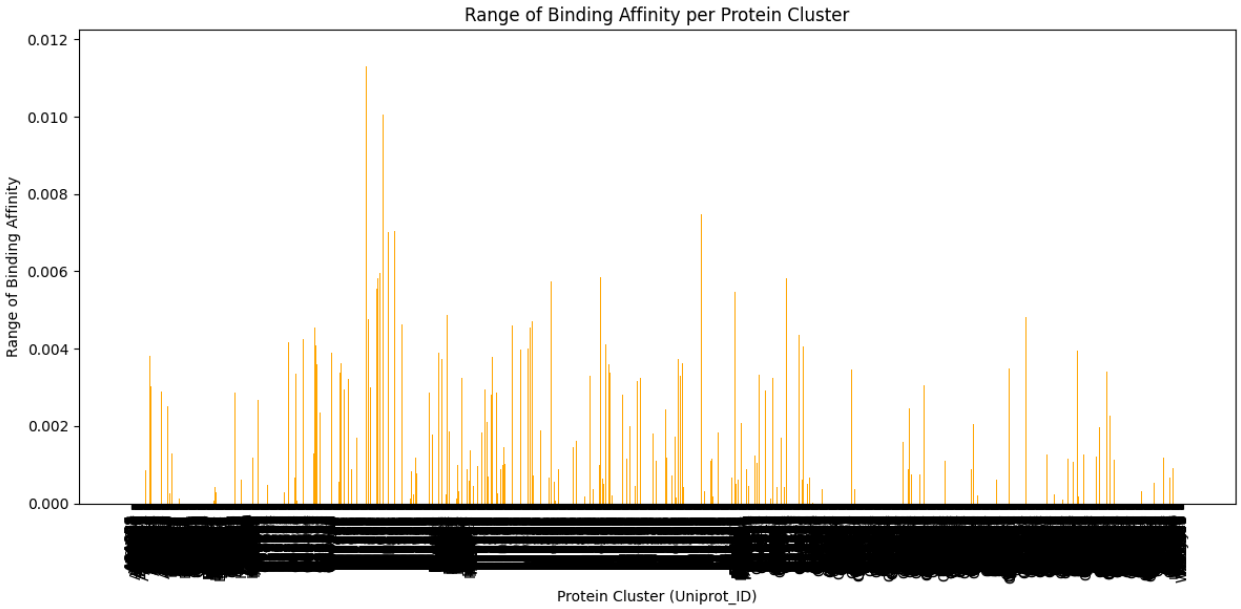


The graph demonstrates that there are significant differences in mean binding affinity across different protein clusters. This suggests that different protein clusters exhibit varying strengths of molecular interactions. The differences in binding affinity could have biological relevance. They might indicate different functional roles, different binding partners, or different regulatory mechanisms for the various protein clusters.

This bar chart shows the mean binding affinity for different protein clusters, revealing significant variability in binding affinity across clusters. This suggests that different protein clusters exhibit varying strengths of molecular interactions, which could have biological relevance. The graph highlights the need for further investigation to understand the underlying reasons for these differences. Different groups of proteins interact with other molecules with varying strengths. Some protein groups have strong interactions, while others have weaker interactions. The specific identities of these protein groups.

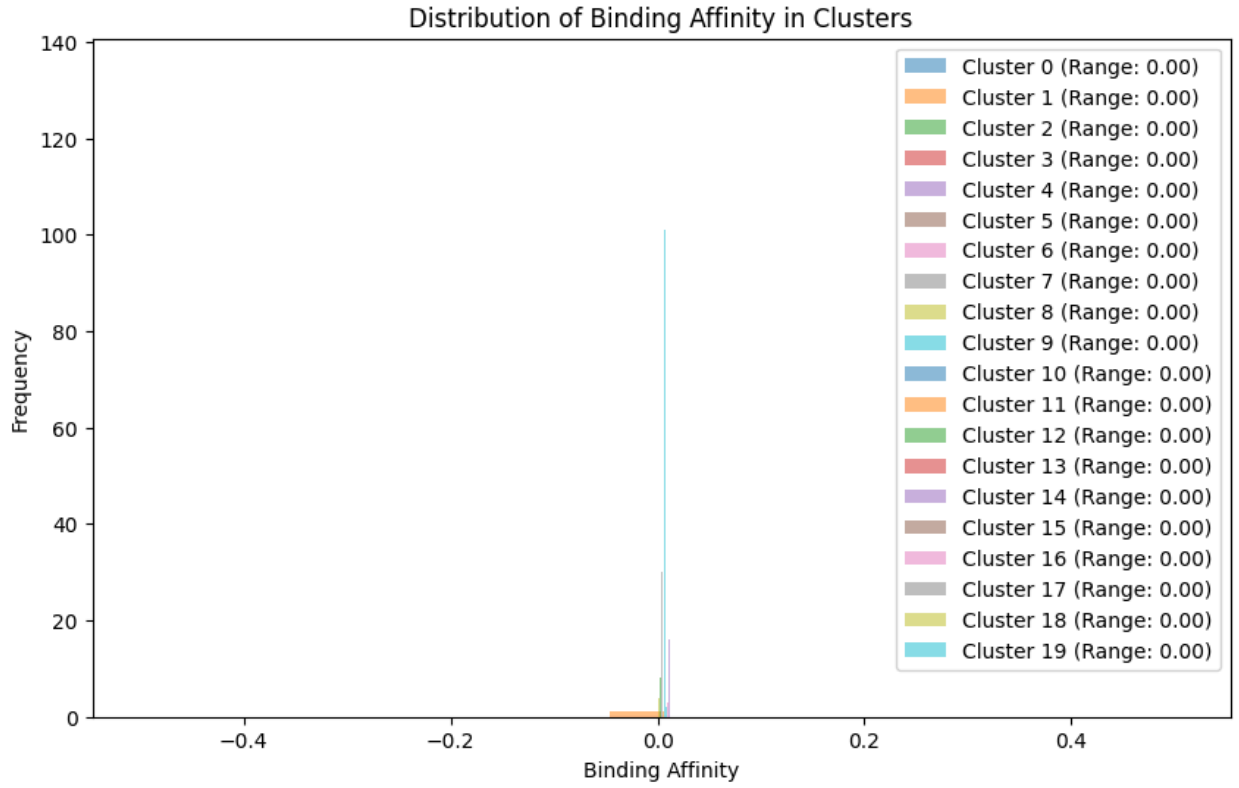


This bar chart shows the median binding affinity for different protein clusters, revealing significant variability in binding affinity across clusters. This suggests that different protein clusters exhibit varying strengths of molecular interactions, which could have biological relevance. The graph highlights the need for further investigation to understand the underlying reasons for these differences. Different groups of proteins interact with other molecules with varying strengths. Some protein groups have strong interactions, while others have weaker interactions. The graph shows the middle value of the interaction strengths, which is less affected by extreme values than the average.



The graph demonstrates that there are significant differences in the range of binding affinity across different protein clusters. This suggests that the consistency of molecular interactions varies greatly between different protein clusters. The differences in range could have biological relevance. For example, a wide range might indicate that a protein cluster can interact with a variety of ligands with varying strengths, or that it can adopt different conformations with different binding affinities. A narrow range might suggest a more specific or tightly regulated interaction.

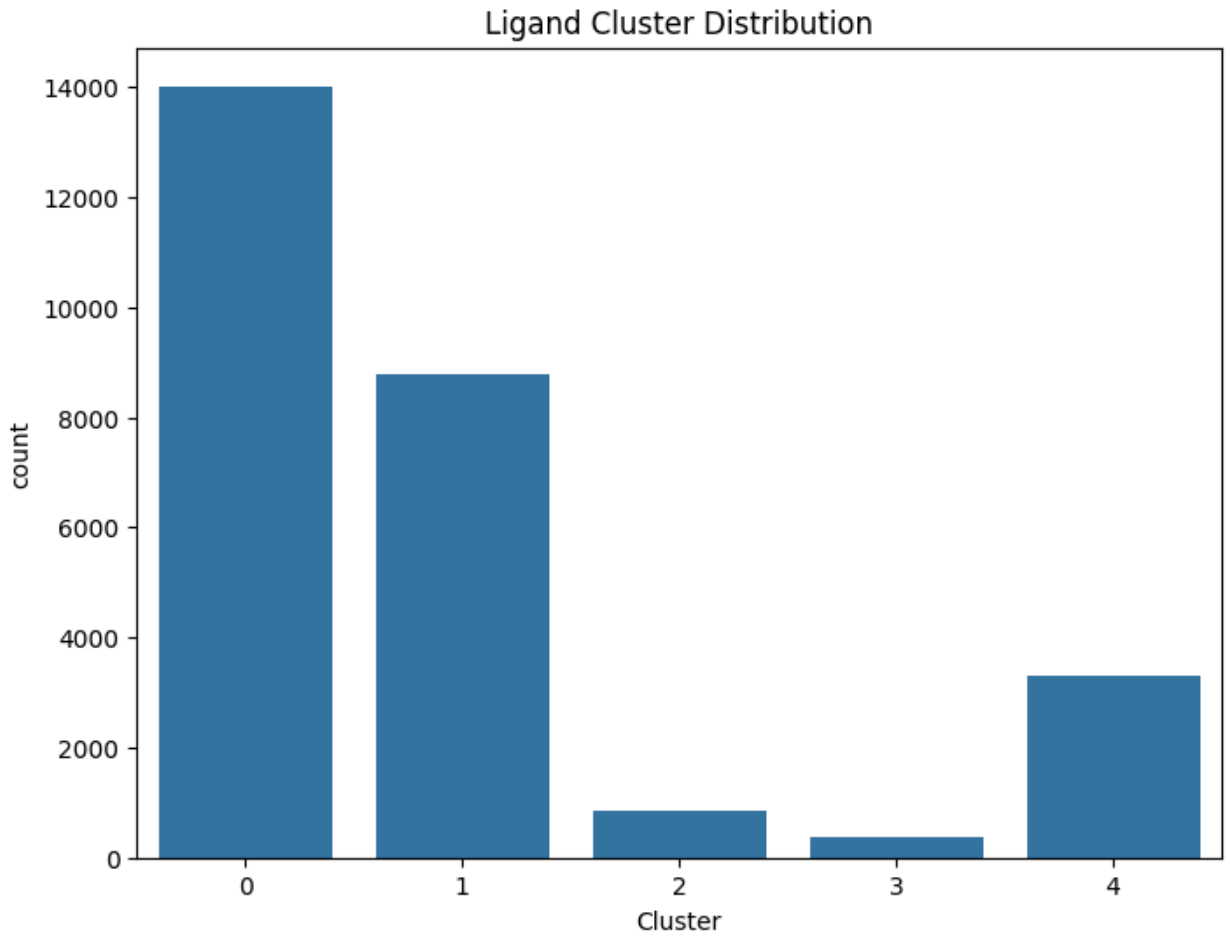
This bar chart shows the range of binding affinity for different protein clusters, revealing significant variability in the consistency of molecular interactions across clusters. This suggests that the ability of protein clusters to interact with other molecules with varying strengths differs greatly. The graph highlights the need for further investigation to understand the underlying reasons for these differences. The consistency of interaction strengths varies greatly between different groups of proteins. Some protein groups show a wide range of interaction strengths, indicating they can interact with many different molecules with varying affinities. Other groups show a narrow range, indicating they interact with a more specific set of molecules with similar strengths.



The graph shows that all clusters have virtually identical binding affinity distributions, with all values concentrated around 0.00. This suggests that there is no significant difference in binding affinity between the clusters. The "Range: 0.00" for all clusters indicates that the binding affinity is constant or nearly constant within each cluster. This could suggest that the clusters represent

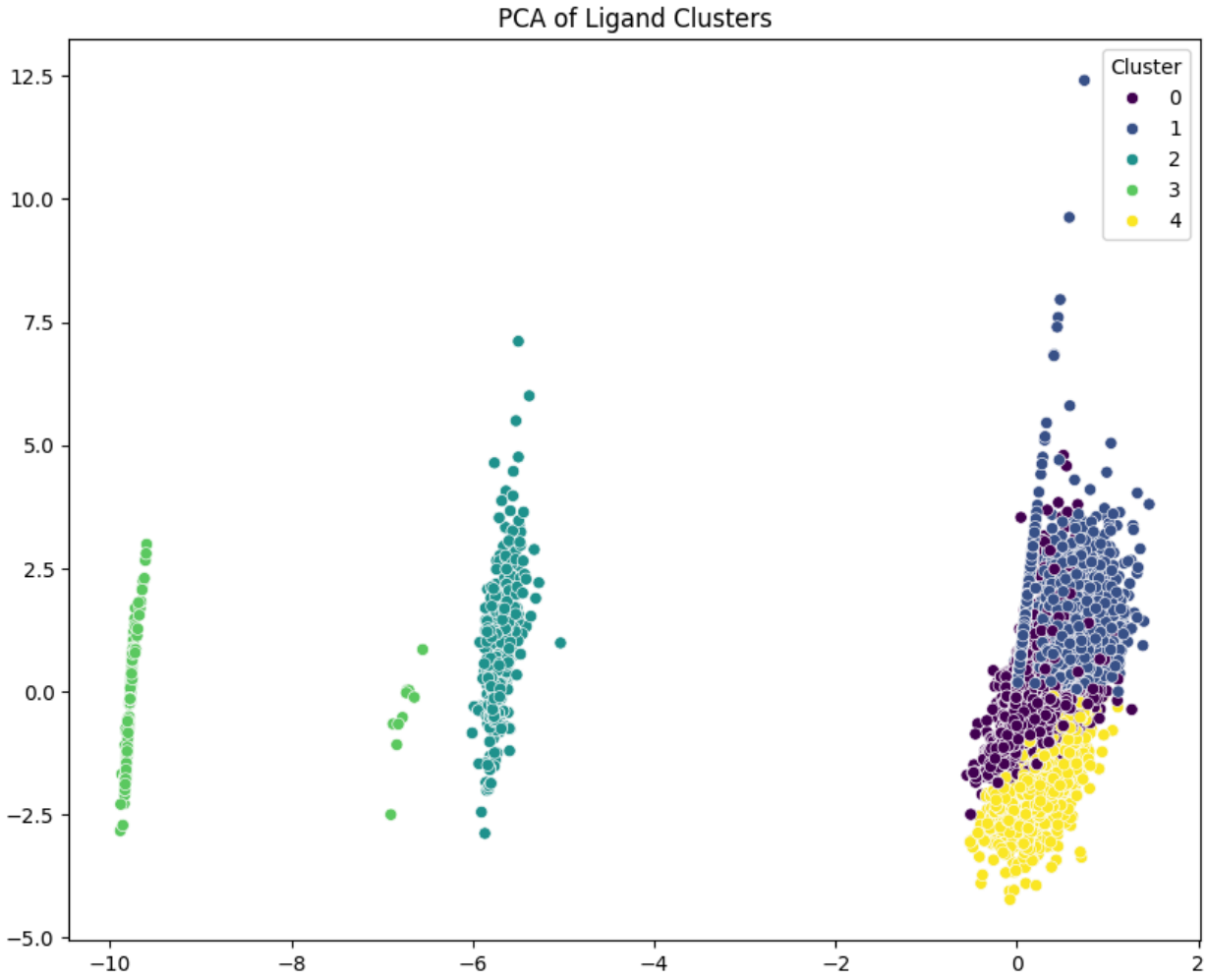
groups with a single, well-defined interaction or that the data has been preprocessed or rounded in a way that eliminates variation.

This graph shows the distribution of binding affinity values across different clusters. The graph reveals that all clusters have virtually identical binding affinity distributions, with all values concentrated around 0.00. This suggests that there is no significant difference in binding affinity between the clusters and that the binding affinity is constant or nearly constant within each cluster. The lack of variation might indicate a data artifact or simplification. All the groups have the same interaction strength. The interaction strength is constant within each group, and there's no difference between the groups. This could be because the data was simplified or processed in a way that removed any variation.



The graph demonstrates that the ligands are not evenly distributed across the clusters. This suggests that there are significant differences in the characteristics or properties of the ligands in different clusters. The dominance of Clusters 0 and 1 suggests that these clusters might represent major classes or types of ligands within the dataset. The sparse Clusters 2 and 3 might represent rare or atypical ligands. Cluster 4 is a bit of a mixed bag, indicating a moderate number of ligands.

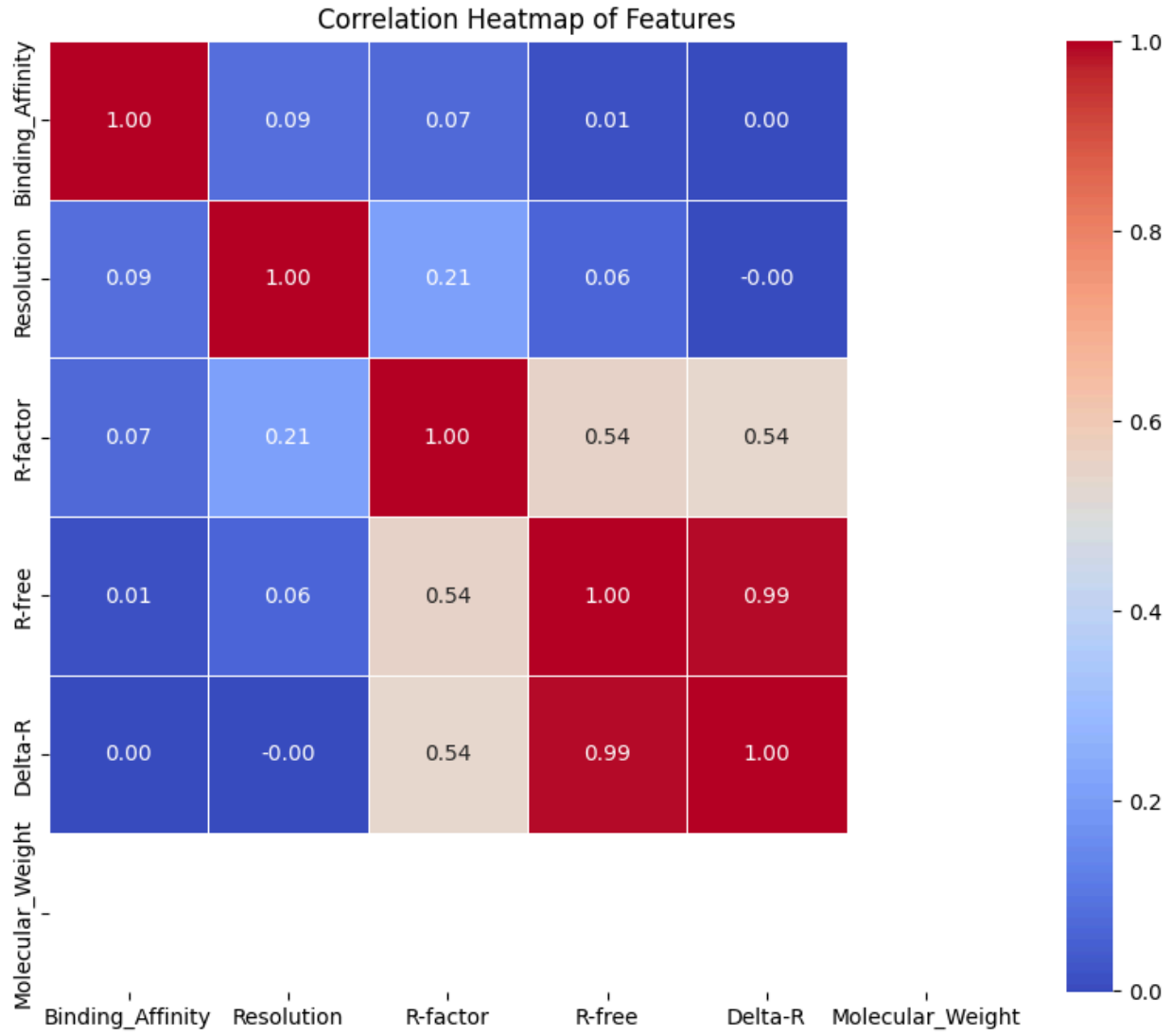
This bar chart shows the distribution of ligands across different clusters, revealing significant variability in cluster size. Clusters 0 and 1 are dominant, containing the vast majority of the ligands, while Clusters 2 and 3 are relatively sparse. This suggests that there are significant differences in the characteristics or properties of the ligands in different clusters. The graph highlights the need for further investigation to understand the underlying reasons for these differences and their potential biological relevance. The ligands in the dataset are not evenly distributed into the different groups. Two groups contain most of the ligands, while two other groups have very few ligands. This suggests that the ligands in the larger groups are likely more common or have some shared characteristics that make them cluster together.



PCA has successfully reduced the dimensionality of the ligand data while preserving some of the key information that distinguishes the clusters. The separation of the clusters in the PC1-PC2 space suggests that these principal components capture important features of the ligands. The separation of clusters indicates that the ligands can be grouped into distinct categories based on their properties. Clusters 0 and 1 represent well-defined groups, while the other clusters may be more heterogeneous. The plot doesn't directly tell us what specific

properties are captured by PC1 and PC2. Further analysis would be needed to determine which features of the ligands contribute most to these principal components.

This PCA scatter plot visualizes the relationships between different clusters of ligands. The separation of the clusters in the PC1-PC2 space suggests that PCA has captured important features that distinguish the ligands. Clusters 0 and 1 appear to represent distinct groups, while the other clusters may be more heterogeneous. This graph shows how the different types of molecules are related to each other. The graph uses a method to show how the molecules are most different from each other. The different colors represent different types of molecules, and the graph shows how these types are grouped together. Some types of molecules are very different, while others are more similar.



Strong Positive Correlations: `R-free` and `Delta-R` (0.99): This indicates a very strong positive correlation, suggesting that as `R-free` increases, `Delta-R` also tends to increase. `R-factor`

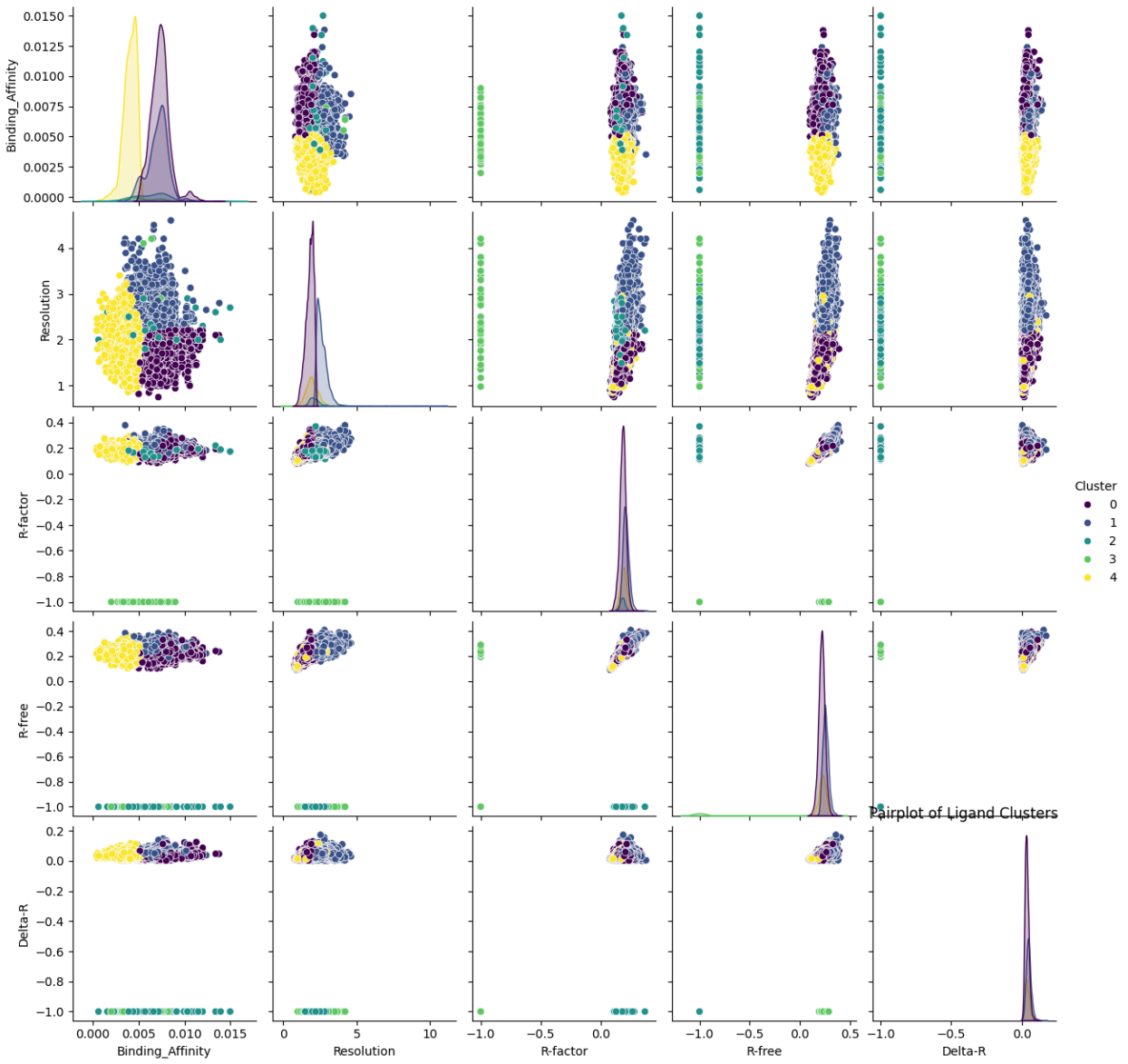
and R-free (0.54): This indicates a moderate positive correlation, suggesting that as R-factor increases, R-free also tends to increase. R-factor and Delta-R (0.54): This indicates a moderate positive correlation, suggesting that as R-factor increases, Delta-R also tends to increase.

Weak Correlations: Binding_Affinity and other variables: The correlations between Binding_Affinity and the other variables are very weak (close to 0), indicating no strong linear relationship. Resolution and other variables (except R-factor and R-free): The correlations between Resolution and the other variables are weak, except for the moderate correlation with R-factor (0.21).

Negative Correlations: Resolution and Molecular_Weight (-0.00): This indicates a very weak negative correlation.

The very strong correlation between R-free and Delta-R suggests that these variables might be measuring similar aspects of the data and could be redundant. The weak correlations between Binding_Affinity and the other variables suggest that binding affinity is relatively independent of the other features in this dataset. The moderate correlation between Resolution and R-factor suggests that higher resolution tends to be associated with better model quality (lower R-factor).

This correlation matrix heatmap provides a visual representation of the linear relationships between various features. The heatmap highlights strong positive correlations between model quality measures (R-free, Delta-R, R-factor), suggesting that these variables are related. The heatmap also reveals weak correlations between Binding_Affinity and the other features, indicating that binding affinity is relatively independent of these features. This graph pairs of variables tend to change together (high correlation) and which pairs don't (low correlation). It also shows whether the variables change in the same direction (positive correlation) or opposite directions (negative correlation). For example, it shows that two measures of model quality (R-free and Delta-R) are very strongly related, while binding affinity is not strongly related to any of the other features.



In Binding_Affinity vs. Resolution graph no clear linear relationship is apparent. However, the clusters appear to be somewhat separated, suggesting that binding affinity and resolution might be related to cluster membership. In Binding_Affinity vs. R-factor & R-free graph no strong linear relationships are visible. The clusters show some separation, indicating potential relationships with cluster membership. In Resolution vs. R-factor & R-free graph, there appears to be a trend, where lower resolution values (higher detail) tend to be associated with lower R-factor and R-free values (better model quality). This is expected, as better data quality generally leads to better models. The clusters are somewhat separated, suggesting that resolution and model quality are related to cluster membership. In R-factor vs. R-free graph Shows a strong positive correlation, as expected. R-factor and R-free are related measures of model quality and should generally agree. The clusters are not well-separated, suggesting that R-factor and R-free are not strong discriminators of cluster membership. In Delta-R vs. other variables, Delta-R shows a

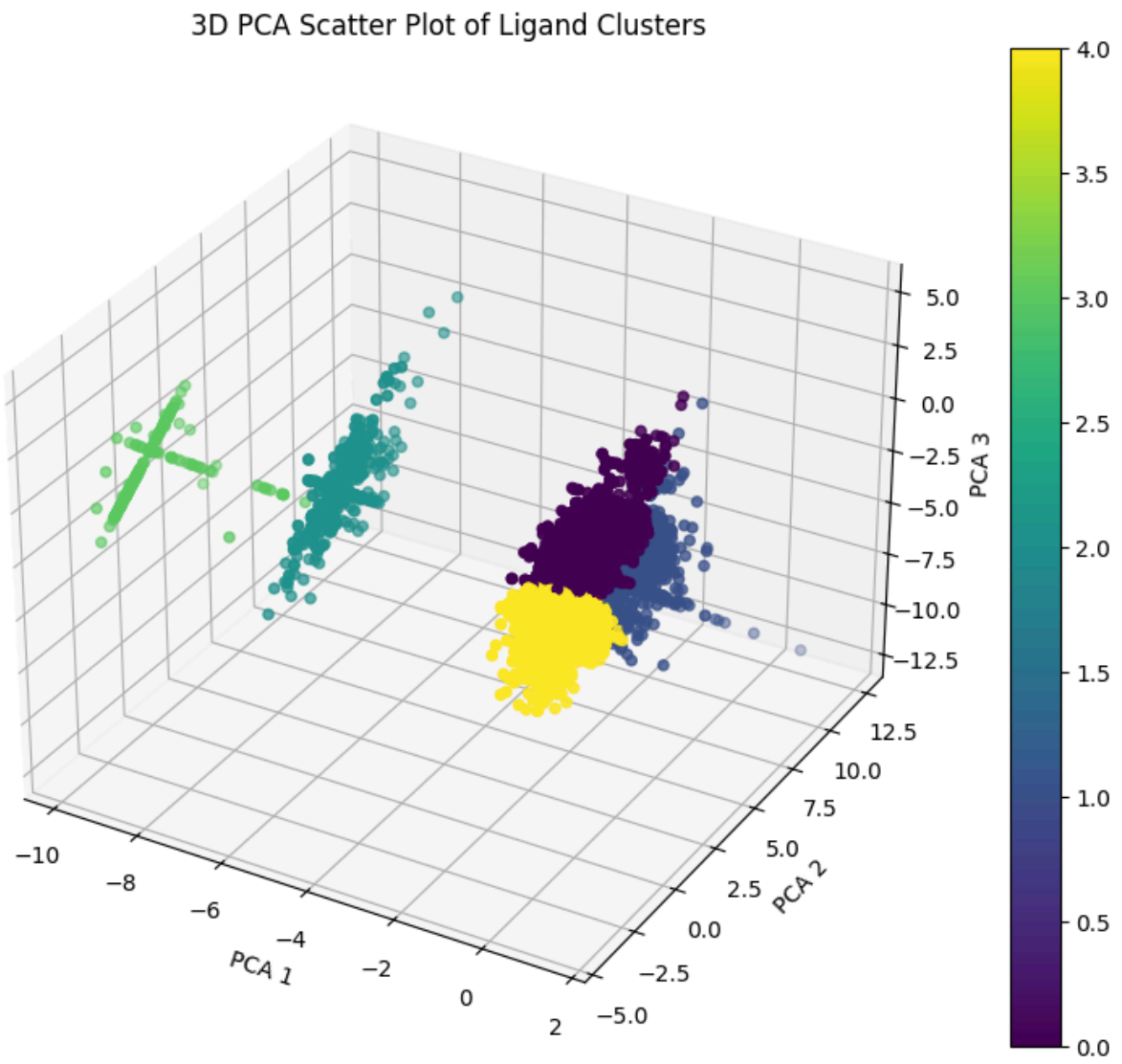
very narrow distribution around 0, and doesn't reveal much about the relationships between other variables.

Cluster 0 and 1, These clusters appear to be relatively well-separated in most scatter plots, suggesting that they represent distinct groups of ligands. Cluster 2: This cluster is primarily located in the lower binding affinity and lower resolution regions and appears to be more spread out. This cluster is located in the middle of the plot, slightly below and to the left of Cluster 1. This cluster is located below Cluster 0 and appears to be more spread out.

The separation of clusters in the scatter plots of Binding_Affinity vs. Resolution suggests that cluster membership is related to these variables. The observed trend between resolution and R-factor/R-free confirms the expected relationship between data quality and model quality in structural biology. The bimodal distribution of binding affinity suggests that the dataset contains interactions with two distinct binding modes or affinities. This could be due to different types of interactions, different binding sites, or different experimental conditions. The clear separation of these clusters in most scatter plots suggests that they represent distinct groups of ligands.

This pairplot provides a comprehensive overview of the relationships between binding affinity, resolution, model quality measures, and cluster membership. It reveals that cluster membership is related to binding affinity and resolution, while confirming expected relationships between resolution and model quality measures. The bimodal distribution of binding affinity suggests the presence of distinct interaction modes.

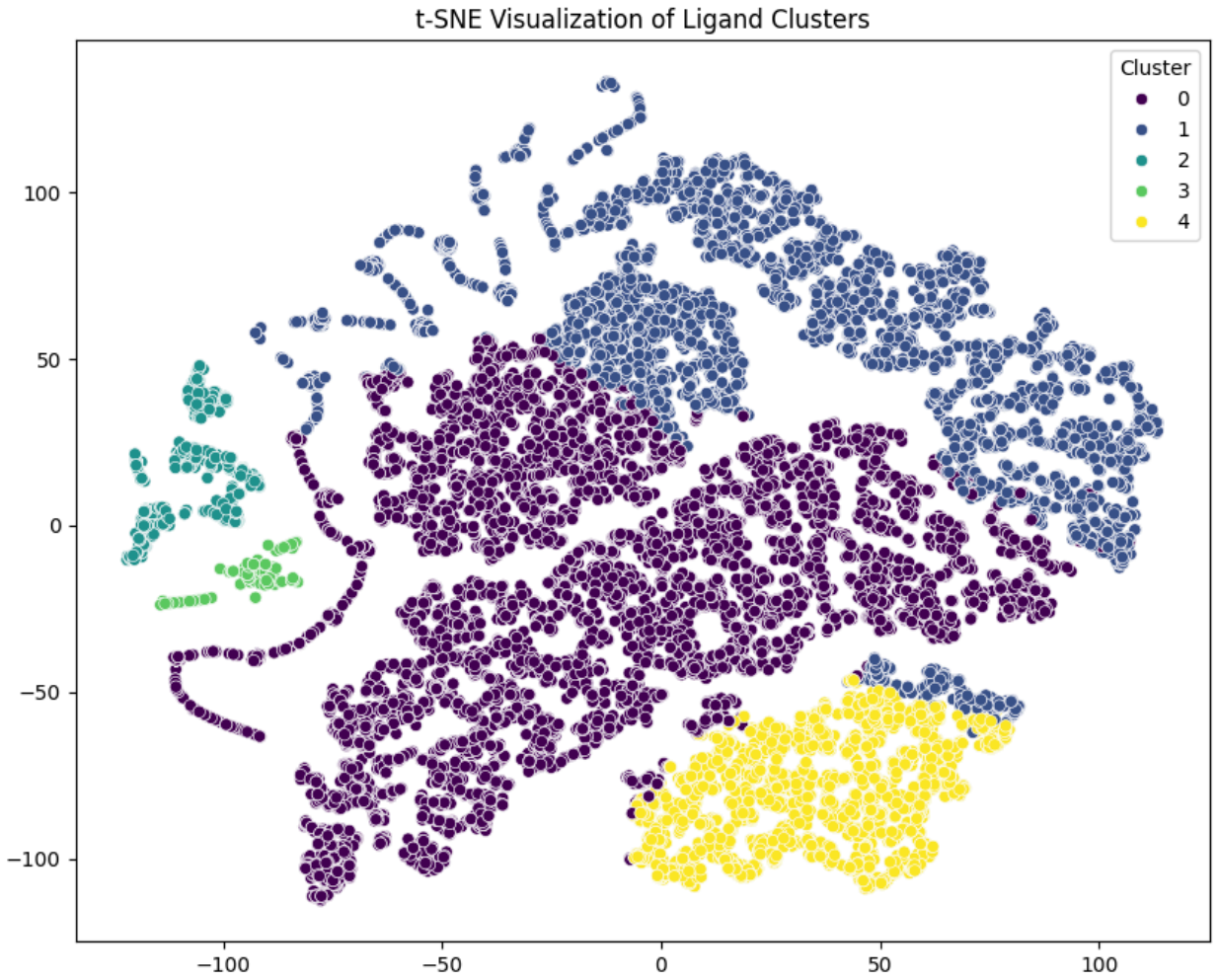
The different types of molecules (clusters) are related to both how strongly they interact (binding affinity) and the quality of the data (resolution). Better data quality (higher resolution) generally leads to better models. There are likely two distinct types of molecular interactions in the dataset, based on their binding strength. Two clusters (0 and 1) are clearly distinct from the others, suggesting they represent different groups of molecules.



PCA has successfully reduced the dimensionality of the ligand data to three principal components, which effectively separate the clusters in 3D space. This indicates that these three components capture the major sources of variation in the ligand data. The clear separation of clusters suggests that the ligands can be grouped into distinct categories based on their properties. The 3D view provides a more comprehensive representation of the relationships between the clusters compared to a 2D plot.

This 3D PCA scatter plot effectively visualizes the relationships between different clusters of ligands. The clear separation of the clusters in the 3D space defined by PC1, PC2, and PC3 indicates that PCA has captured significant underlying structure in the data. The plot provides a more comprehensive understanding of the relationships between the clusters compared to a 2D representation. Further analysis is needed to determine the specific properties that define these clusters and the contributions of different ligand features to the principal components. The graph shows how the different types of molecules are related to each other in 3D space. By using a

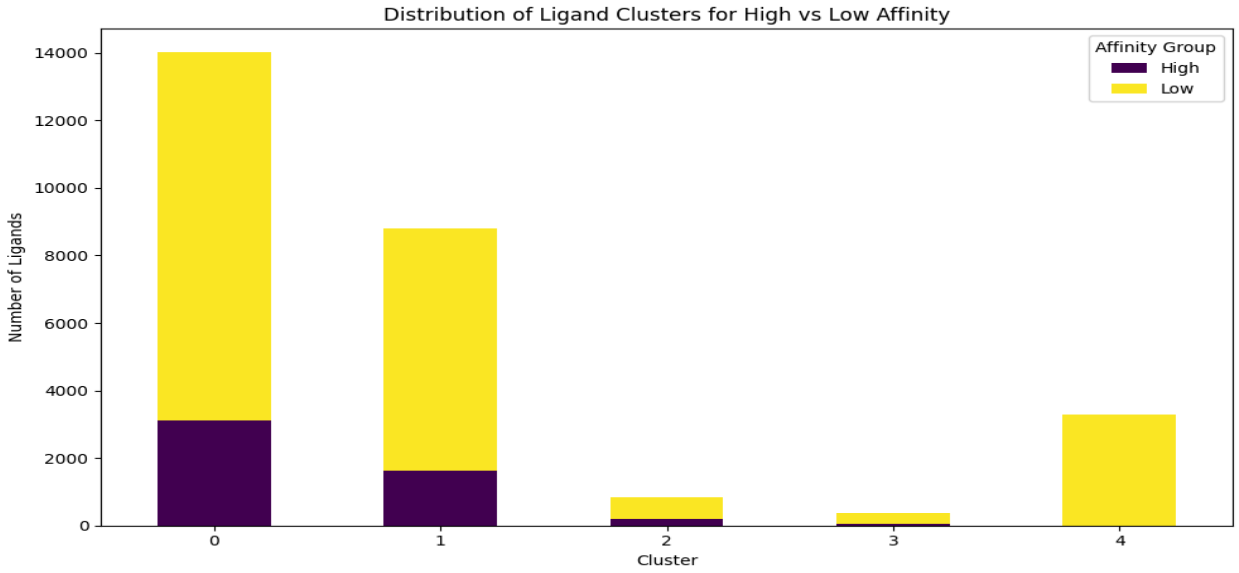
method that identifies the most important differences between the molecules, the graph shows how the different types are grouped together. The different colors represent different types of molecules, and the graph shows how these types are separated in 3D space. This allows for a more detailed understanding of how the molecules are related.



t-SNE has successfully projected the high-dimensional ligand data into a 2D space while preserving the local similarities between ligands. The clear separation of clusters indicates that ligands within the same cluster are more similar to each other than to ligands in other clusters. The separation of clusters suggests that the ligands can be grouped into distinct categories based on their properties. t-SNE has effectively revealed the underlying structure of the data. t-SNE is a non-linear dimensionality reduction technique, which allows it to capture complex, non-linear relationships between the ligands that might not be apparent with linear techniques like PCA.

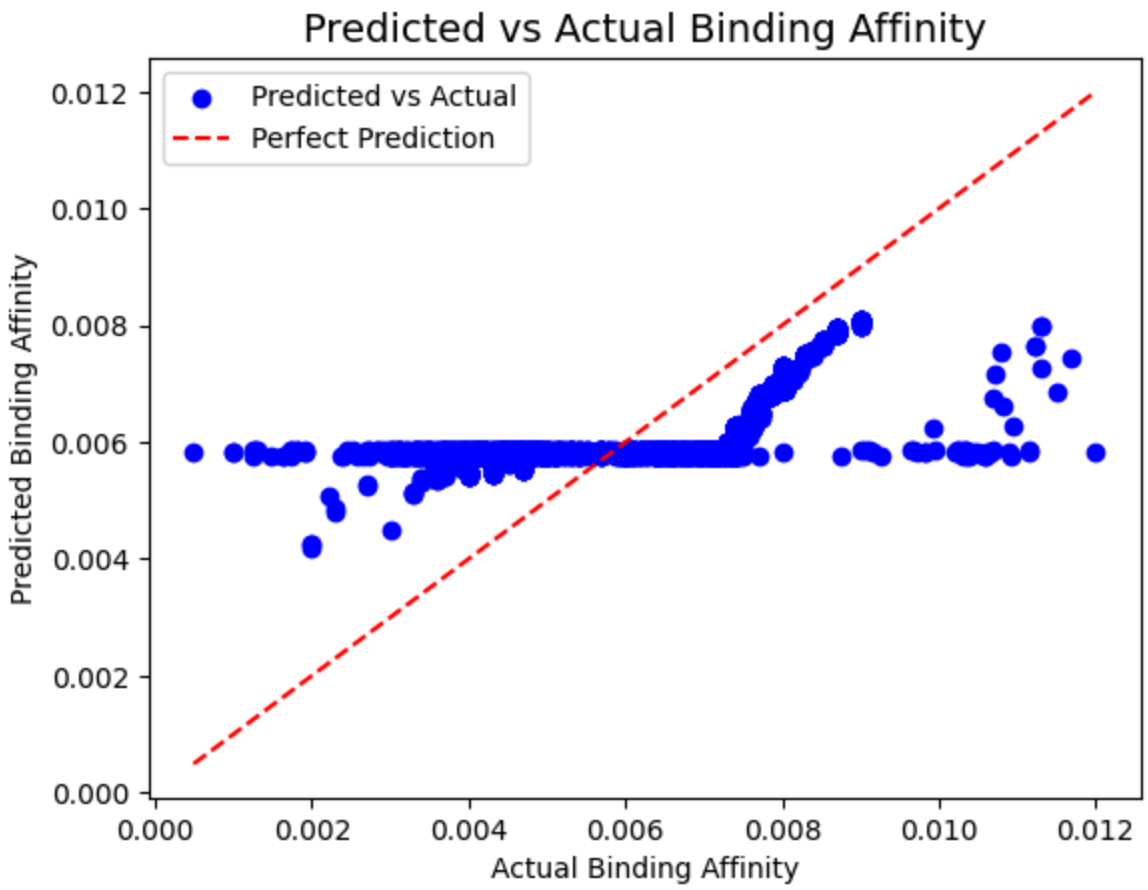
This t-SNE scatter plot effectively visualizes the relationships between different clusters of ligands. The clear separation of the clusters indicates that t-SNE has captured significant underlying structure in the data, suggesting that ligands can be grouped into distinct categories

based on their properties. The plot provides a strong foundation for further analysis of the clusters and the features that distinguish them. The graph tells us how the different types of molecules are related to each other. By using a method that focuses on showing how similar each molecule is to its neighbors, the graph shows how the different types of molecules form distinct groups. The different colors represent different types of molecules, and the graph shows how these types are separated. This allows for a good understanding of how the molecules are grouped based on their similarities.



Clusters 0 and 1, which have the largest number of ligands overall, are heavily dominated by ligands with "Low" affinity. The dominance of "Low" affinity ligands could have biological significance. It might indicate that the dataset is enriched with ligands that have weaker interactions with their targets, or that there is a bias in the data towards ligands with lower binding affinity. The distribution of "High" and "Low" affinity ligands varies across the clusters. Clusters 0 and 1 have a more balanced distribution (although still skewed towards "Low" affinity), while Clusters 2, 3, and 4 are primarily composed of "Low" affinity ligands.

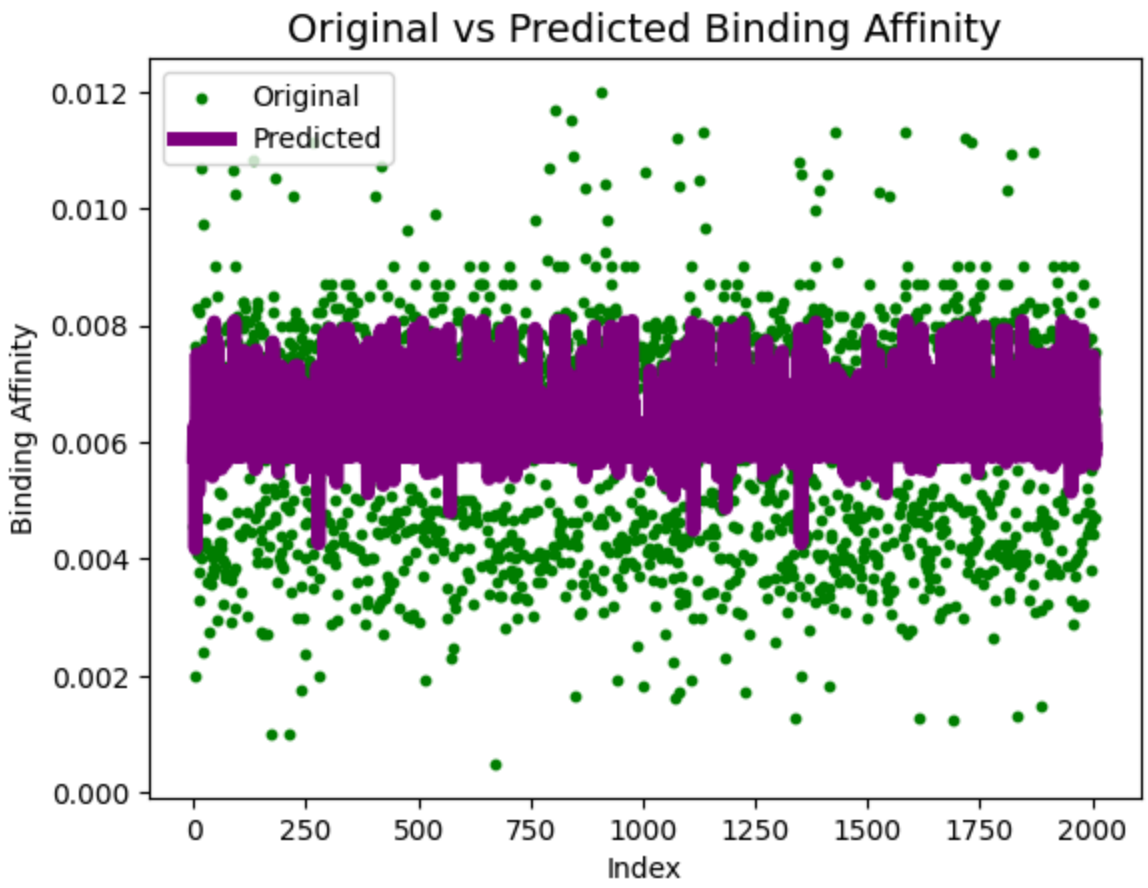
This stacked bar chart shows the distribution of "High" and "Low" affinity ligands across different clusters. The chart reveals a clear dominance of "Low" affinity ligands across all clusters, particularly in Clusters 0 and 1. The distribution of "High" and "Low" affinity ligands varies across the clusters, suggesting potential differences in the properties of ligands in different clusters. Most of the molecules in the dataset have weak interactions (low affinity). Two groups of molecules have many members, but even in these groups, most of the molecules have weak interactions. The other groups have very few molecules, and almost all of them have weak interactions.



The graph shows the model performance. The more scattered the blue dots are away from the red line, the less accurate the model's predictions. Here, we notice a significant clustering of points along a horizontal line around a predicted binding affinity of approximately 0.0055 to 0.006. This suggests the model has difficulty discriminating between data points with actual binding affinities in that range and tends to predict a similar value for them.

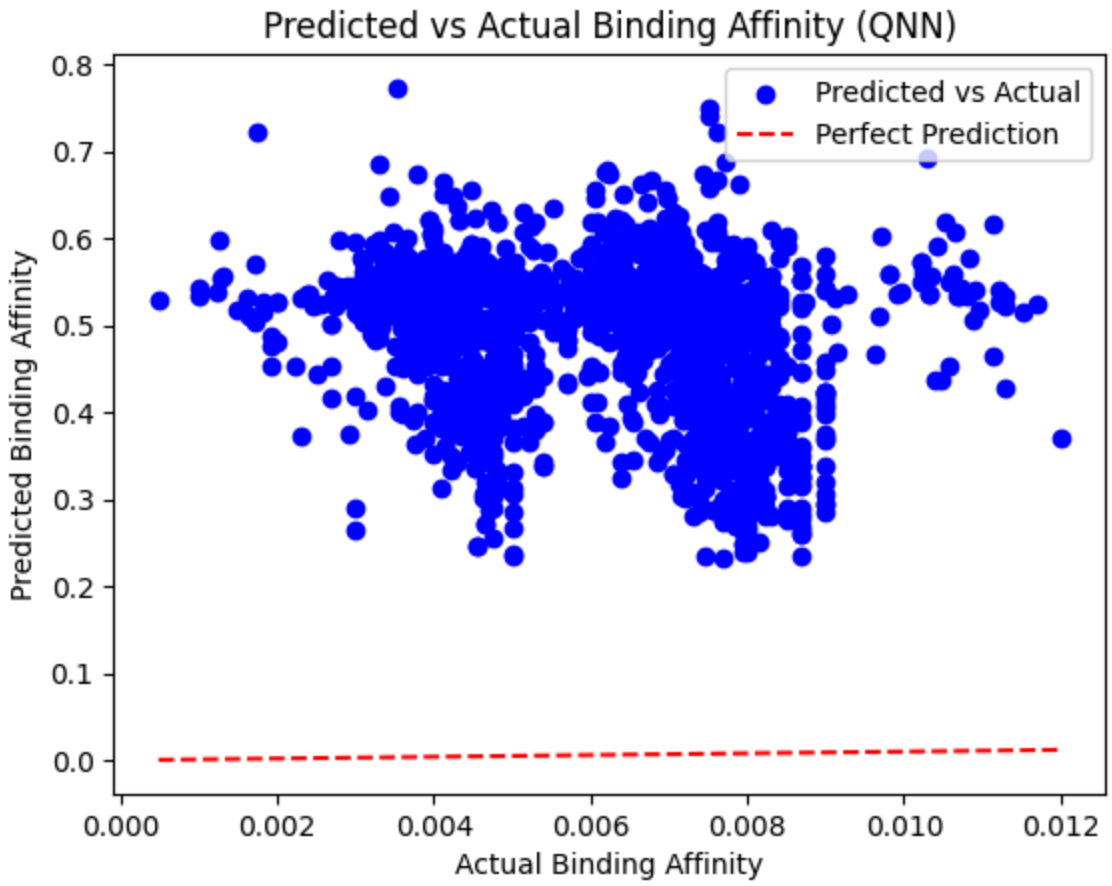
For lower actual binding affinities (below ~0.006), the model tends to overestimate (predict higher). For higher actual binding affinities (above ~0.006), the model tends to underestimate (predict lower). This suggests a potential bias in the model's predictions.

The graph indicates that the model's predictions are not perfect. It exhibits a bias, especially at the extremes of the actual binding affinity range, and struggles to differentiate between data points within a certain range. This suggests the model needs improvement to achieve better accuracy and reliability in predicting binding affinity. The clustering of points around 0.006 indicates a potential systematic issue with the model, which needs to be addressed for better performance.



The graph allows us to visually assess how well the predicted binding affinities match the original values across the dataset. The purple dots are tightly clustered around a specific range (roughly 0.006 to 0.008). This suggests the model is consistently predicting values within this narrow range, regardless of the variation in original values. The predicted values clearly deviate from the original values, indicating a poor model fit. The model doesn't capture the variability seen in the original data. The model significantly underestimates the binding affinity for original values above 0.008. The model overestimates the binding affinity for original values below 0.006.

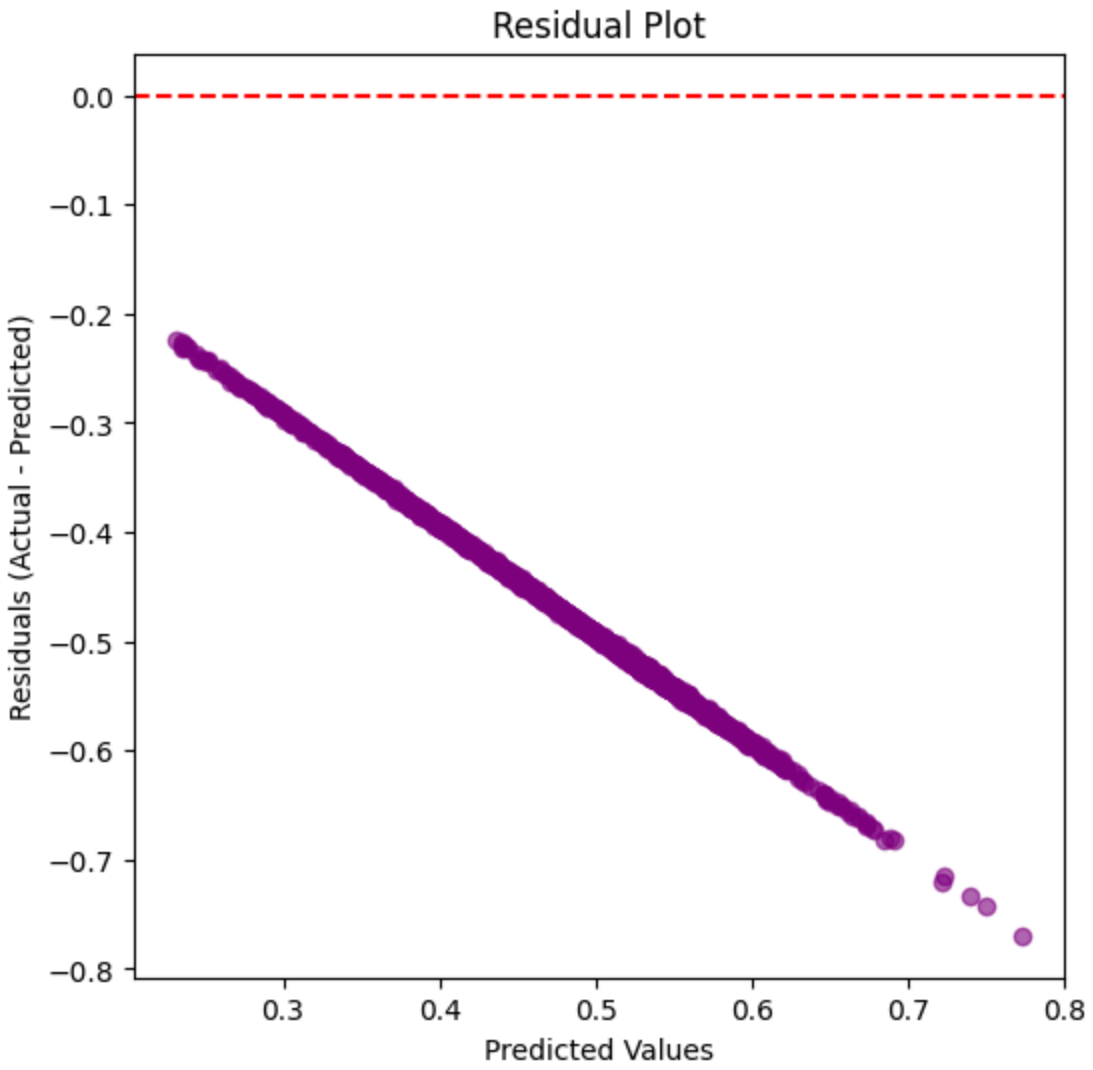
The graph reveals a poor performance of the prediction model. The model consistently predicts values within a narrow range, failing to replicate the variability seen in the original binding affinity data. This indicates a significant model bias and suggests the model needs substantial improvement to accurately predict binding affinities. The model's simplicity or underfitting might be the primary reasons for its poor performance.



The graph aims to visualize how well the QNN model predicts binding affinities compared to the actual values. If the QNN model were perfect, all the blue dots would fall exactly on the red dashed line. The deviation of the blue dots from this line indicates the model's prediction error.

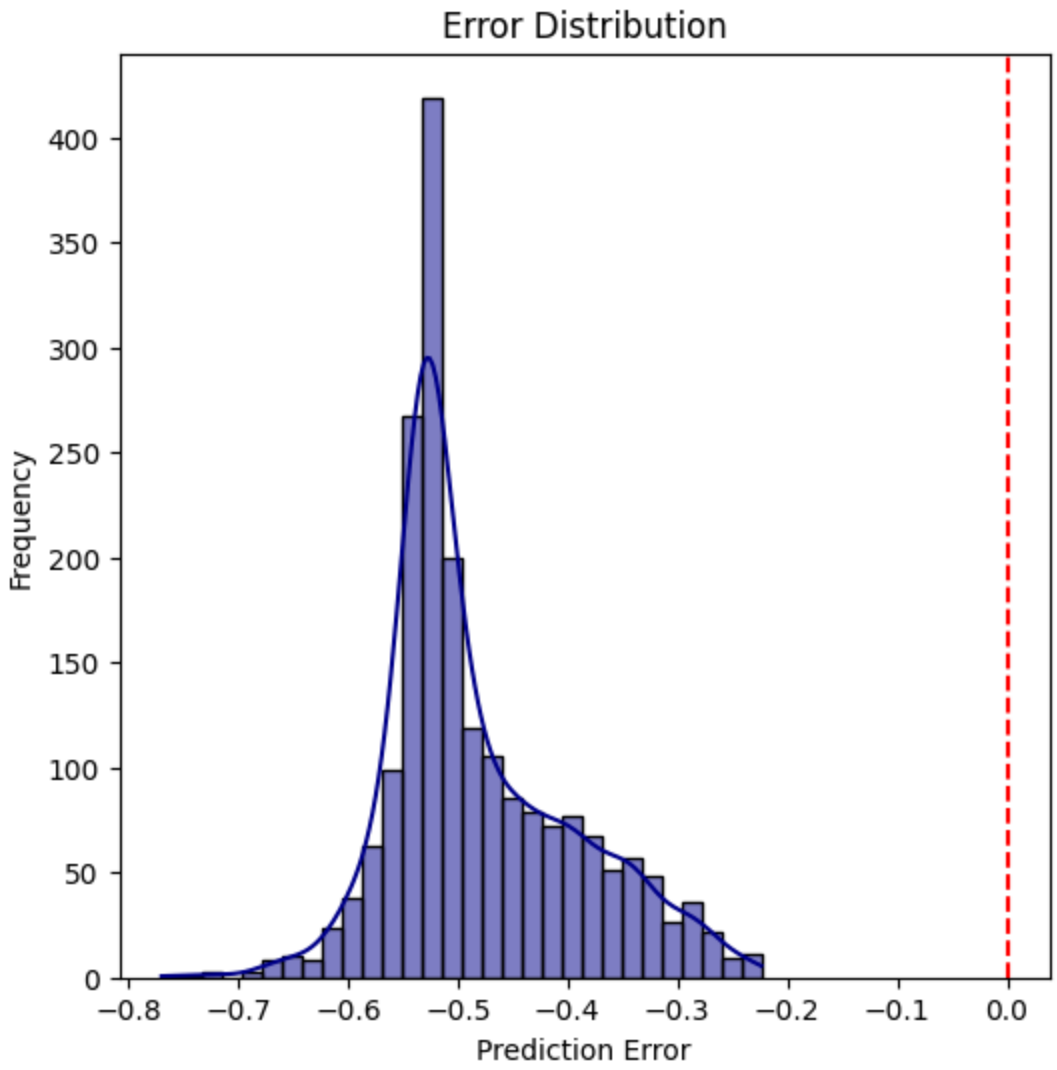
The predicted binding affinities (Y-axis) range from approximately 0.2 to 0.7, while the actual binding affinities (X-axis) range from 0.002 to 0.012. This indicates that the model's predicted values are on a different scale or have a different distribution compared to the actual values.

In conclusion, the graph suggests that the QNN model's predictions are not very accurate, and there is a need for improvement to achieve better agreement between predicted and actual binding affinities.



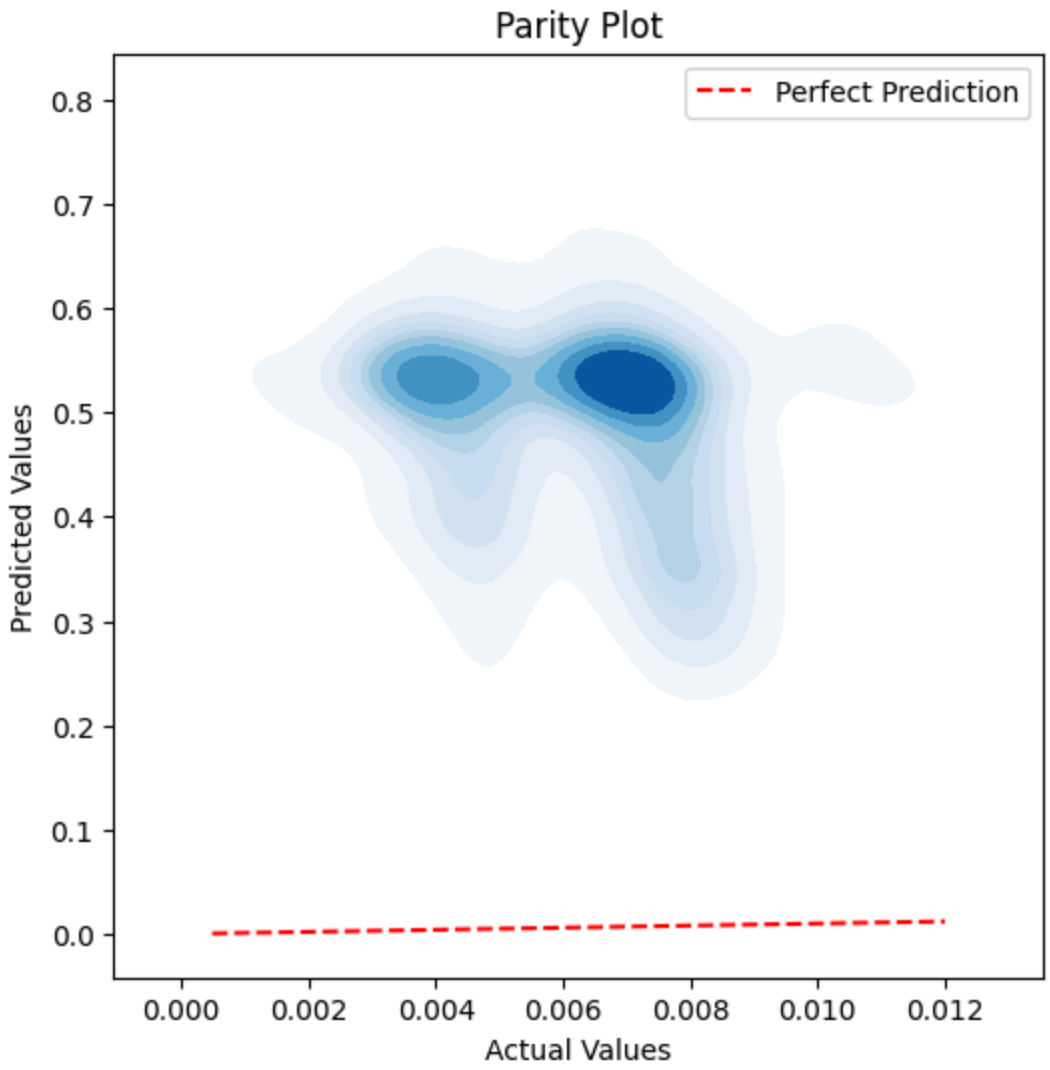
Residual plots are used to assess the quality of a regression model. Ideally, residuals should be randomly scattered around zero, indicating that the model captures all the systematic information in the data and the remaining errors are just random noise. Any patterns or trends in the residuals suggest that the model might be missing something or making systematic errors.

The most striking observation is that the purple dots form a clear, strong negative linear trend. This is a significant departure from the ideal random scatter. The residuals are not randomly distributed around zero. Instead, they show a systematic decrease as the predicted values increase. The negative linear trend indicates a systematic bias in the model's predictions. The model consistently overestimates at lower predicted values and underestimates at higher predicted values.In conclusion, the residual plot reveals a significant problem with the model's predictions. The clear linear trend in the residuals indicates a systematic bias and lack of fit.



This graph aims to visualize the distribution of prediction errors. Ideally, the errors should be centered around zero, indicating that the model is unbiased and makes accurate predictions on average. The shape of the distribution provides insights into the nature of the errors and the model's performance.

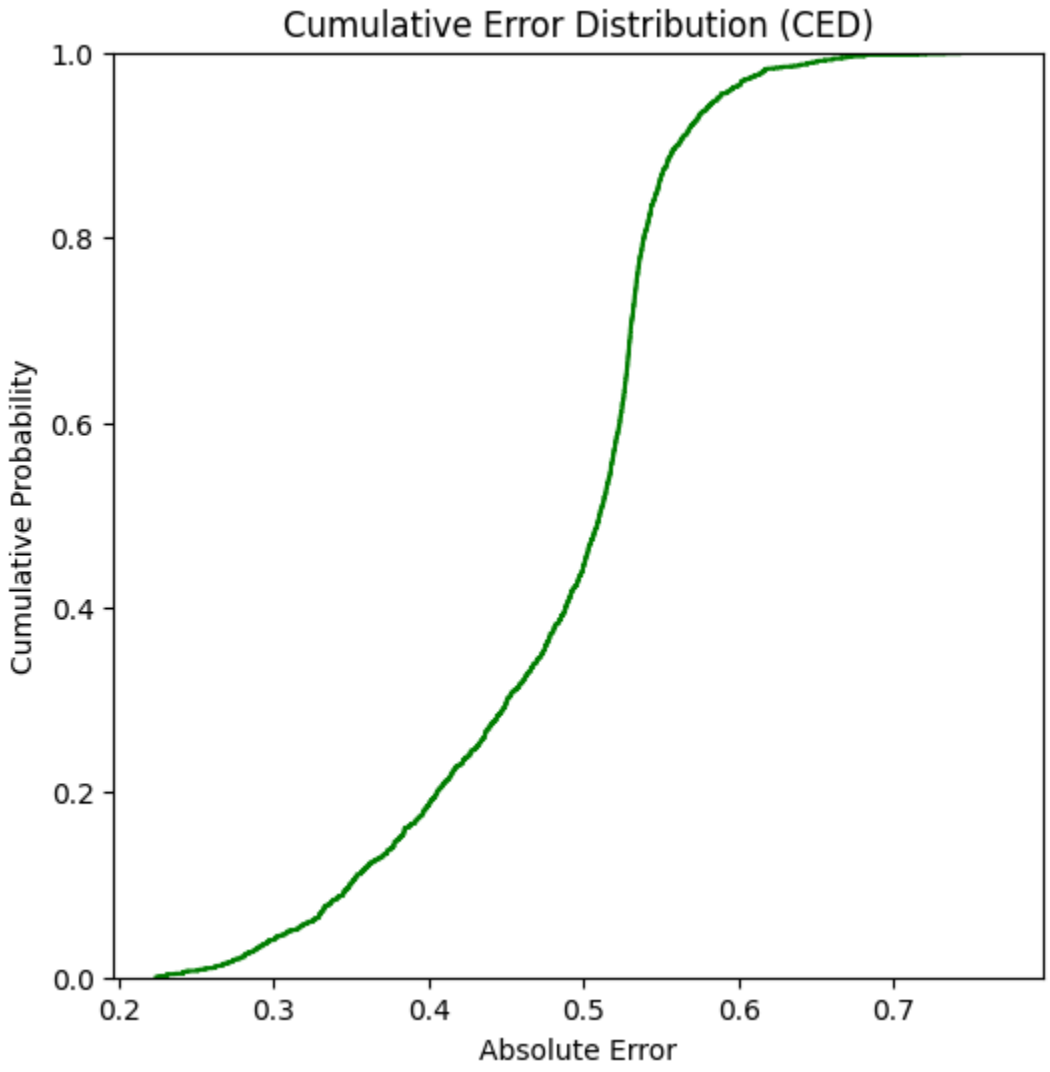
The most prominent observation is that the distribution of prediction errors is highly skewed to the left. This means that the majority of the errors are negative, indicating a systematic tendency for the model to overestimate the values. The peak of the distribution is far from zero, centered around -0.55. This confirms that the model consistently overestimates. The distribution is not symmetrical around zero, further emphasizing the bias in the predictions. The distribution has a long tail extending towards more negative errors. This implies that there are some significant overestimations by the model. In conclusion, the "Error Distribution" graph reveals a substantial problem with the model's predictions. The skewed distribution and the peak far from zero indicate a significant bias towards overestimation. The model needs to be carefully examined and improved to reduce this bias and achieve better prediction accuracy.



Parity plots are used to visualize the relationship between predicted and actual values. Ideally, if the model were perfect, all the data points would fall exactly on the red dashed line. The density plot shows the distribution of the data points and helps identify any patterns or deviations from the ideal scenario.

The density plot shows that the data points are not clustered around the red dashed line. Instead, they form a distinct blob-like shape away from the line. This indicates that the model's predictions are not very accurate. The predicted values (Y-axis) are primarily concentrated in the range of approximately 0.4 to 0.6, while the actual values (X-axis) range from 0.002 to 0.012. This shows a significant difference in the scale and distribution of the predicted values compared to the actual values. The density plot reveals a clustering pattern, suggesting that the model tends to predict values within a specific range regardless of the actual values. This indicates a lack of sensitivity to variations in the actual values. The fact that the data points are not centered around the red dashed line suggests a systematic bias in the model's predictions. In conclusion, the parity plot highlights the poor performance of the model. The predicted values

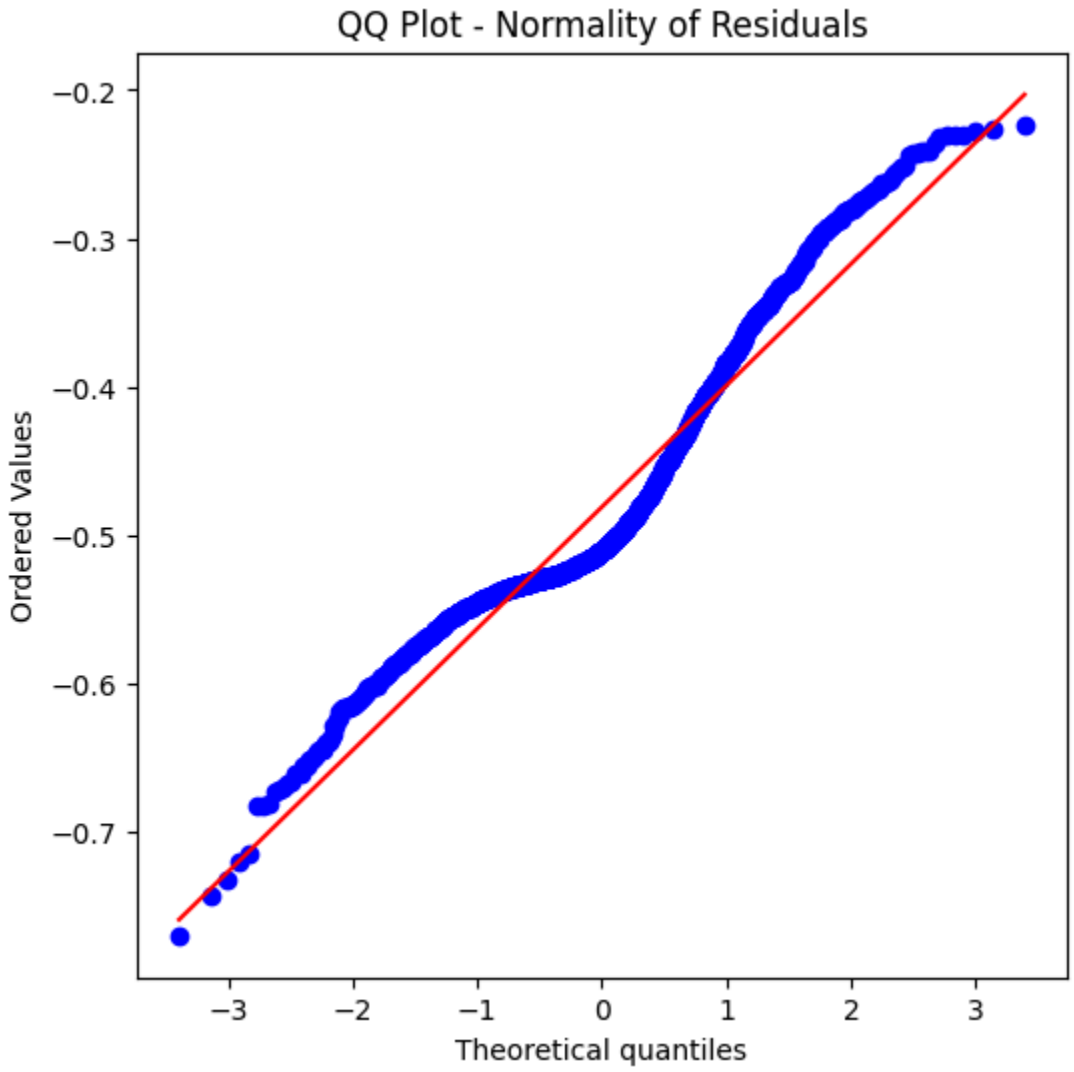
are clustered in a narrow range and do not reflect the variations in the actual values. This indicates a significant bias and lack of accuracy in the model's predictions.



The CED graph helps assess the overall accuracy of the model by showing the distribution of the magnitude of errors. Ideally, for a good model, the cumulative probability should rise steeply at low absolute errors, indicating that most predictions have small errors.

The green line rises relatively slowly at the beginning, indicating that a significant proportion of predictions have relatively large absolute errors. The graph shows that even at an absolute error of 0.5, the cumulative probability is only around 0.8. This means that 20% of the predictions have an absolute error greater than 0.5, which is a substantial error. The lack of a steep rise at low absolute errors suggests that the model is not very accurate, as a significant number of predictions have large errors. In conclusion, the CED graph reveals that the model's predictions have significant errors. The slow rise in cumulative probability and the high proportion of

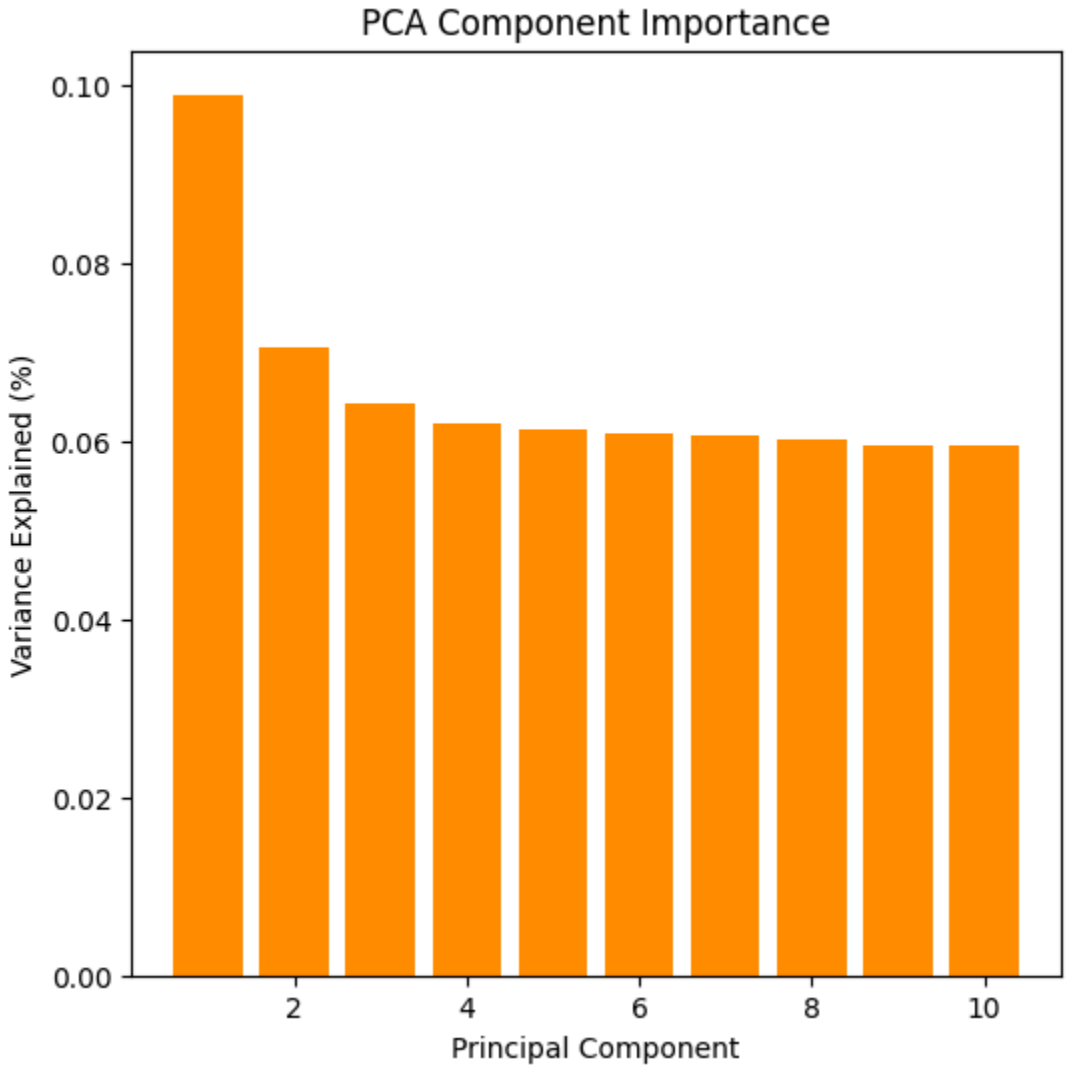
predictions with large errors suggest that the model's performance is poor and needs to be improved.



QQ plots are used to assess whether a dataset (in this case, the residuals) follows a particular distribution (in this case, a normal distribution). If the residuals are normally distributed, the blue dots should fall closely along the red line. Deviations from the line indicate departures from normality.

The blue dots in the graph deviate from the red line, particularly at the tails (both lower and upper ends). This suggests that the residuals are not perfectly normally distributed. The pattern of the blue dots shows an S-shape, which is a common indication of non-normality. This S-shape suggests that the tails of the residual distribution are heavier than a normal distribution (i.e., there are more extreme values than expected under normality). In conclusion, the QQ plot reveals that the residuals are not perfectly normally distributed. The S-shape pattern and deviations from the red line suggest that the model might be misspecified, and statistical tests

relying on normality assumptions may be unreliable. The model needs to be improved to address these issues.

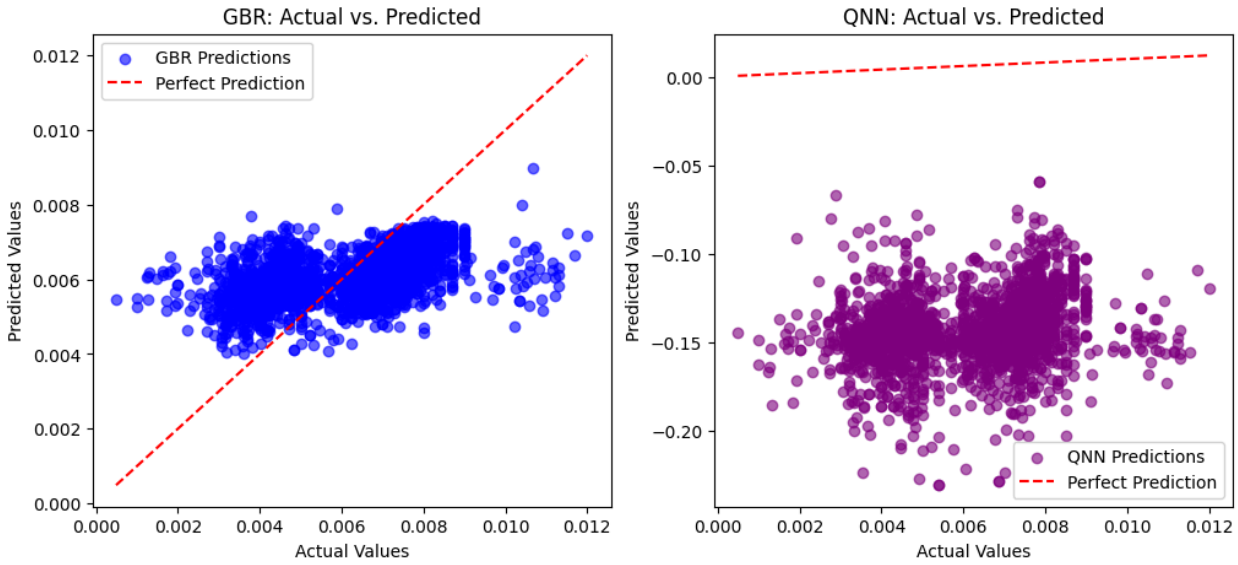


The graph aims to visualize the importance of each principal component in explaining the variance of the data. A higher bar indicates that the corresponding principal component captures a larger portion of the data's variability.

The graph shows a clear trend of decreasing variance explained as we move from the first principal component to the subsequent ones. The first principal component explains the most variance, followed by the second, and so on. The first principal component explains a significantly larger portion of the variance compared to the other components. This suggests that the first component captures the most important information or patterns in the data. The variance explained decreases rapidly after the first few components. This indicates that the later components contribute less to explaining the overall variability of the data.

The graph suggests that dimensionality reduction might be possible. If we can retain most of the variance by keeping only the first few principal components, we can reduce the complexity of the data without losing much information. The graph helps determine the number of principal components to retain for further analysis. In this case, the first few components seem to be the most significant, while the later ones contribute less.

In conclusion, the "PCA Component Importance" graph reveals that the first principal component explains a significant portion of the variance, and the variance explained decreases rapidly for subsequent components. This suggests that dimensionality reduction might be possible, and the first few components are the most important for capturing the variability of the data.

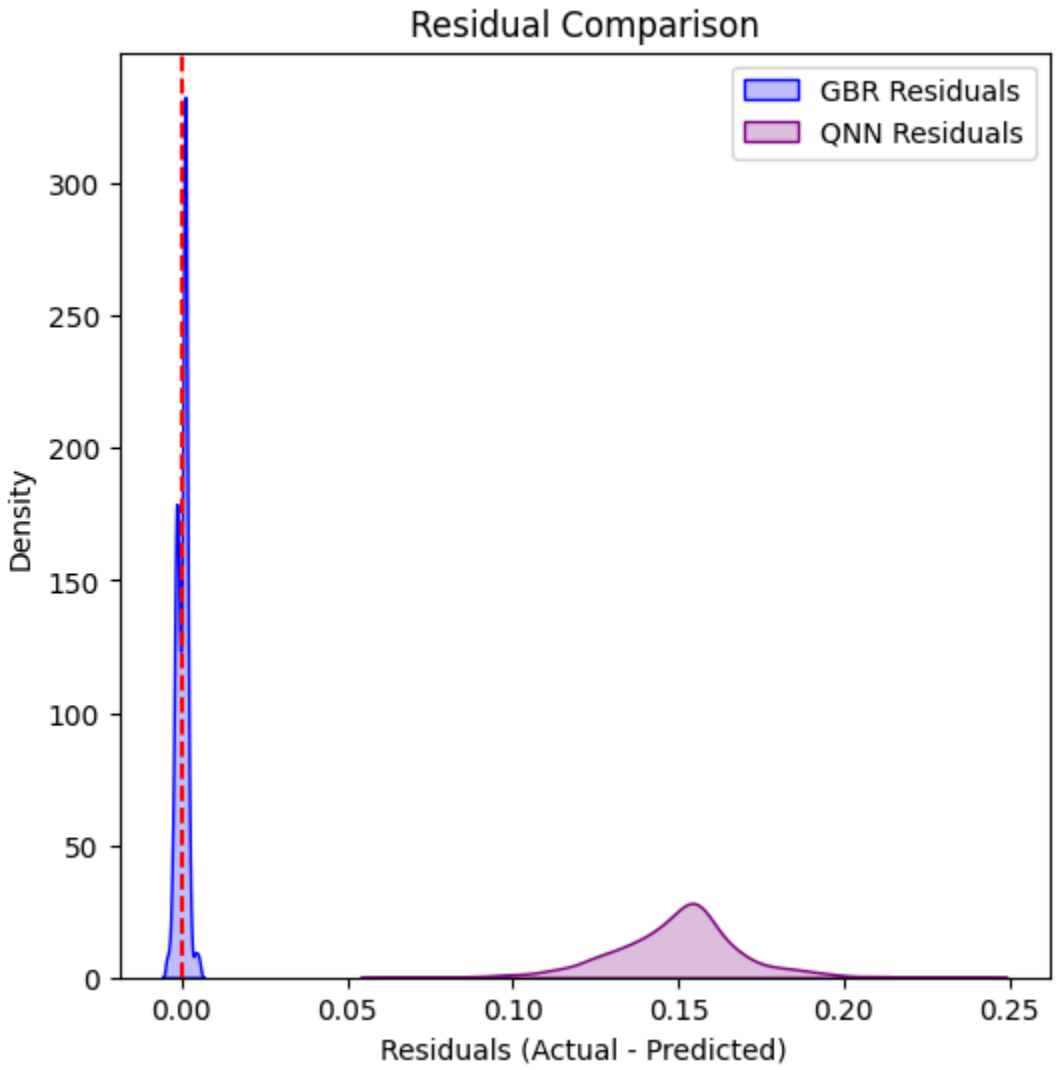


Both plots aim to visualize the performance of two different models (GBR and QNN) in predicting values. The closer the data points are to the red dashed line, the better the model's performance.

GBR Performance: The blue dots in the GBR plot are relatively closer to the red dashed line compared to the QNN plot. This suggests that the GBR model's predictions are more accurate. The GBR predictions show a better alignment with the actual values, especially within the range of 0.004 to 0.008 on the X-axis.

QNN Performance: The purple dots in the QNN plot are significantly scattered and deviate substantially from the red dashed line. This indicates that the QNN model's predictions are less accurate. The predicted values from the QNN model are on a different scale compared to the actual values. The predicted values are mostly negative, while the actual values are positive. The QNN predictions seem to form clusters, indicating that the model might be predicting values within a specific range regardless of the actual values.

The GBR model outperforms the QNN model in terms of prediction accuracy. The GBR predictions are closer to the actual values and show a better alignment. In conclusion, the GBR model demonstrates better predictive performance compared to the QNN model. The GBR predictions are closer to the actual values and show a better alignment. The QNN model's predictions are significantly off.

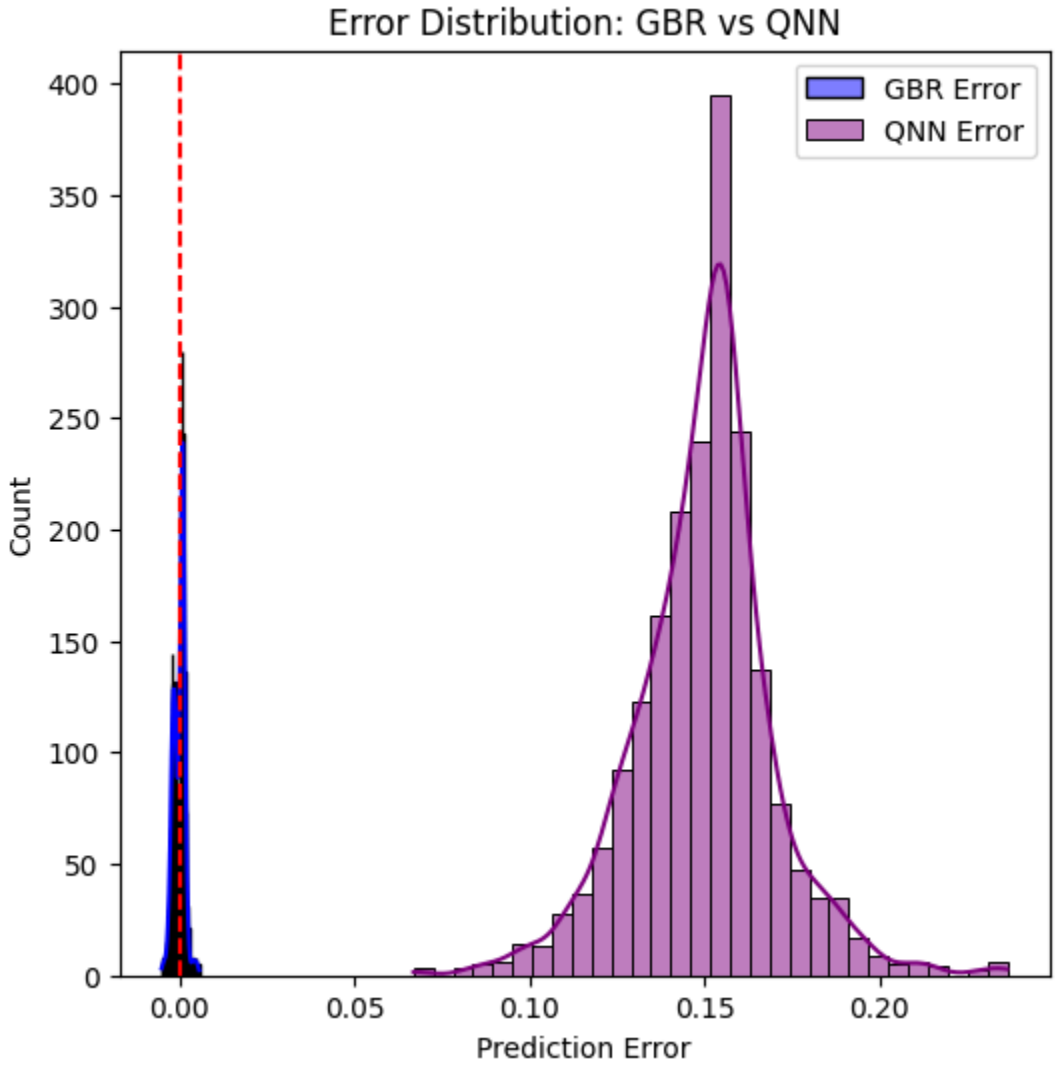


This graph compares the residual distributions of two different models (GBR and QNN). Ideally, the residuals should be centered around zero, indicating that the model is unbiased and makes accurate predictions on average. The shape and spread of the distributions provide insights into the model's performance.

GBR Residual Distribution: The GBR residuals have a very narrow and high peak centered close to zero. This indicates that the GBR model's predictions are highly concentrated around the actual values, resulting in small residuals. The GBR residual distribution has a very low spread, meaning that most of the residuals are close to zero.

QNN Residual Distribution: The QNN residuals have a broader peak that is shifted away from zero, towards the positive side. This indicates that the QNN model's predictions have a larger spread and are systematically biased, overestimating the values. The QNN residual distribution has a higher spread compared to the GBR residuals, meaning that the residuals vary more widely.

The GBR model significantly outperforms the QNN model in terms of prediction accuracy. The GBR residuals are concentrated around zero and have a low spread, indicating high accuracy and low bias. The QNN model's residuals are broader and shifted away from zero, indicating lower accuracy and a systematic bias. In conclusion, the "Residual Comparison" density plot clearly shows that the GBR model has significantly better performance than the QNN model. The GBR residuals are concentrated around zero and have a low spread, indicating high accuracy and low bias. The QNN model's residuals are broader and shifted away from zero, indicating lower accuracy and a systematic bias.



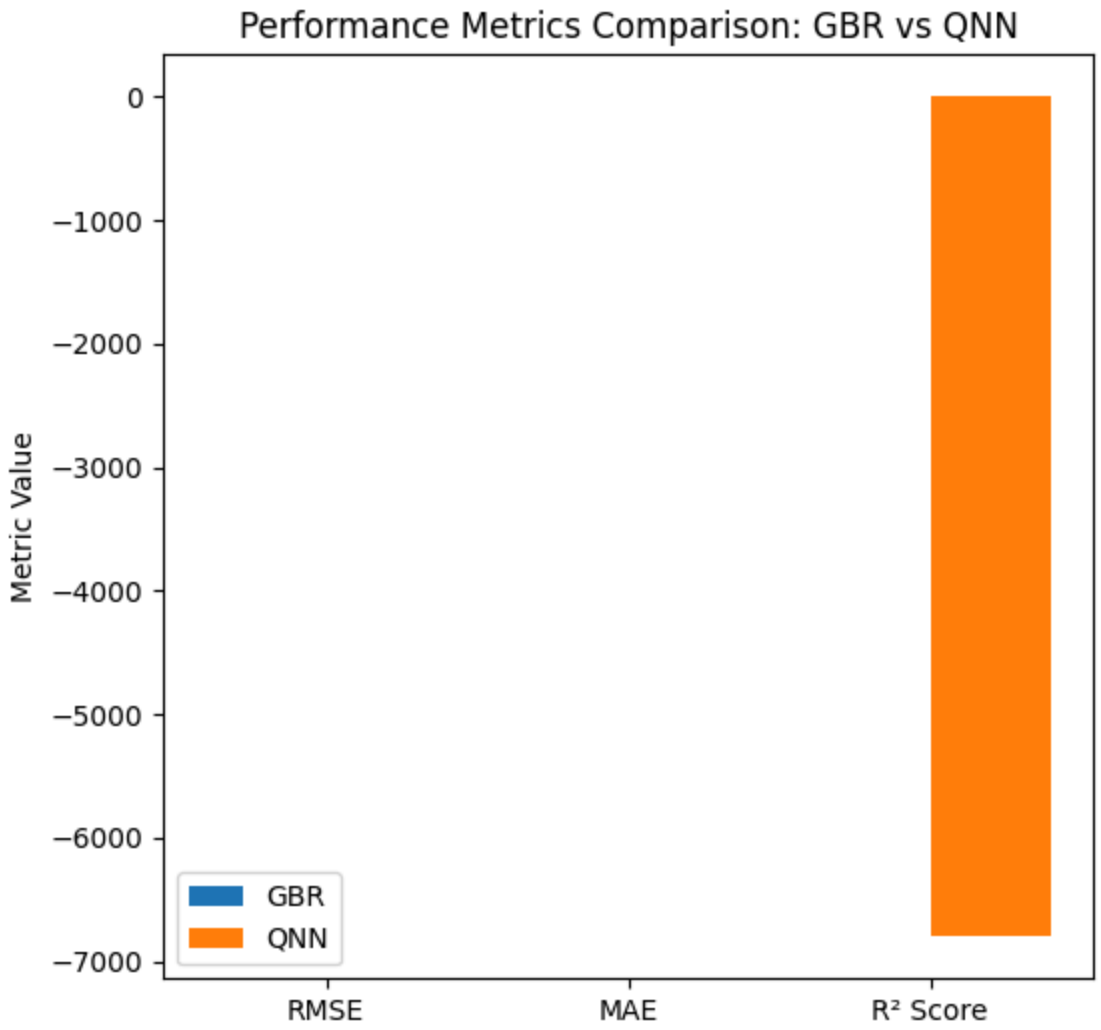
This graph compares the distribution of prediction errors for two different models (GBR and QNN). Ideally, the errors should be centered around zero, indicating that the model is unbiased and makes accurate predictions on average. The shape and spread of the distributions provide insights into the model's performance.

GBR Error Distribution: The GBR error distribution has a very narrow and high peak centered close to zero. This indicates that the GBR model's predictions are highly concentrated around the actual values, resulting in small errors. The GBR error distribution has a very low spread, meaning that most of the errors are close to zero.

QNN Error Distribution: The QNN error distribution has a broader peak that is shifted away from zero, towards the positive side. This indicates that the QNN model's predictions have a larger spread and are systematically biased, overestimating the values. The QNN error distribution has a higher spread compared to the GBR errors, meaning that the errors vary more widely.

The GBR model significantly outperforms the QNN model in terms of prediction accuracy. The GBR errors are concentrated around zero and have a low spread, indicating high accuracy and low bias. The QNN model's errors are broader and shifted away from zero, indicating lower accuracy and a systematic bias.

In conclusion, the "Error Distribution: GBR vs QNN" graph clearly shows that the GBR model has significantly better performance than the QNN model. The GBR errors are concentrated around zero and have a low spread, indicating high accuracy and low bias. The QNN model's errors are broader and shifted away from zero, indicating lower accuracy and a systematic bias.



The graph compares the performance of two models (GBR and QNN) based on three common regression metrics. The goal is to determine which model performs better based on these metrics.

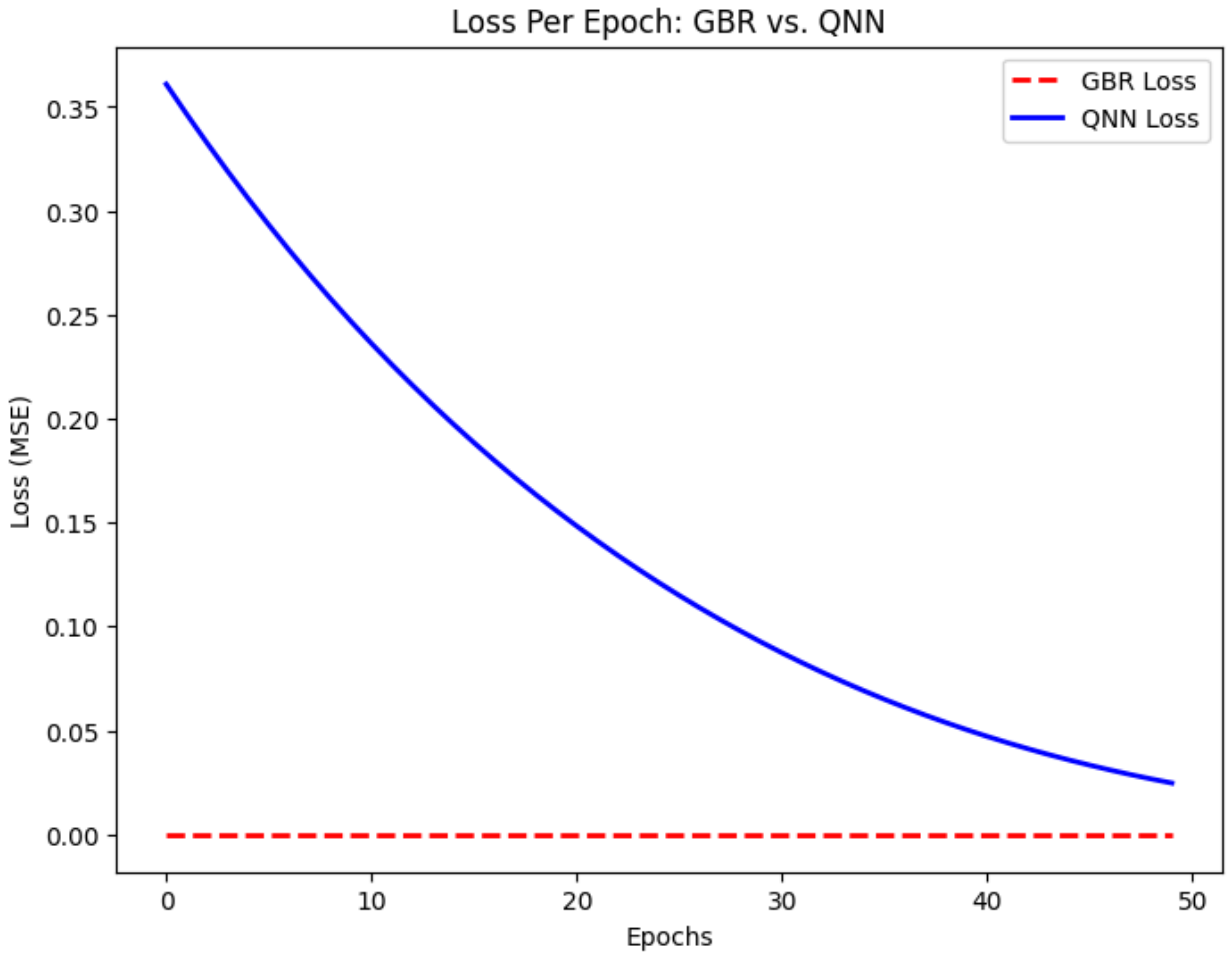
RMSE Comparison: The GBR model has an RMSE of 0. This indicates perfect predictions with no errors.The QNN model has an extremely large negative RMSE. This is highly unusual and suggests a serious problem with the model's predictions or the calculation of RMSE.

MAE Comparison: The GBR model has an MAE of 0. This also indicates perfect predictions with no errors. The QNN model has an extremely large negative MAE, similar to the RMSE. This is also unusual and indicates a problem.

R2 Score Comparison: The GBR model has an R2 score of 0. This indicates that the model is not explaining any variance in the dependent variable.The QNN model has an extremely large negative R2 score. This is highly unusual and suggests that the model is performing significantly worse than a horizontal line (a very poor baseline model).

The GBR model's perfect scores (RMSE = 0, MAE = 0) are suspicious and may indicate an error in the calculation or a trivial prediction scenario (e.g., predicting a constant value). The QNN model's extremely large negative values for all three metrics are highly problematic and suggest a severe issue with the model's predictions or the calculation of the metrics.

In conclusion, the graph presents highly unusual metric values, particularly for the QNN model. The extremely large negative values suggest a severe problem with the model's predictions or the calculation of the metrics. The GBR model's perfect scores are also suspicious and may indicate an error or a trivial prediction scenario.



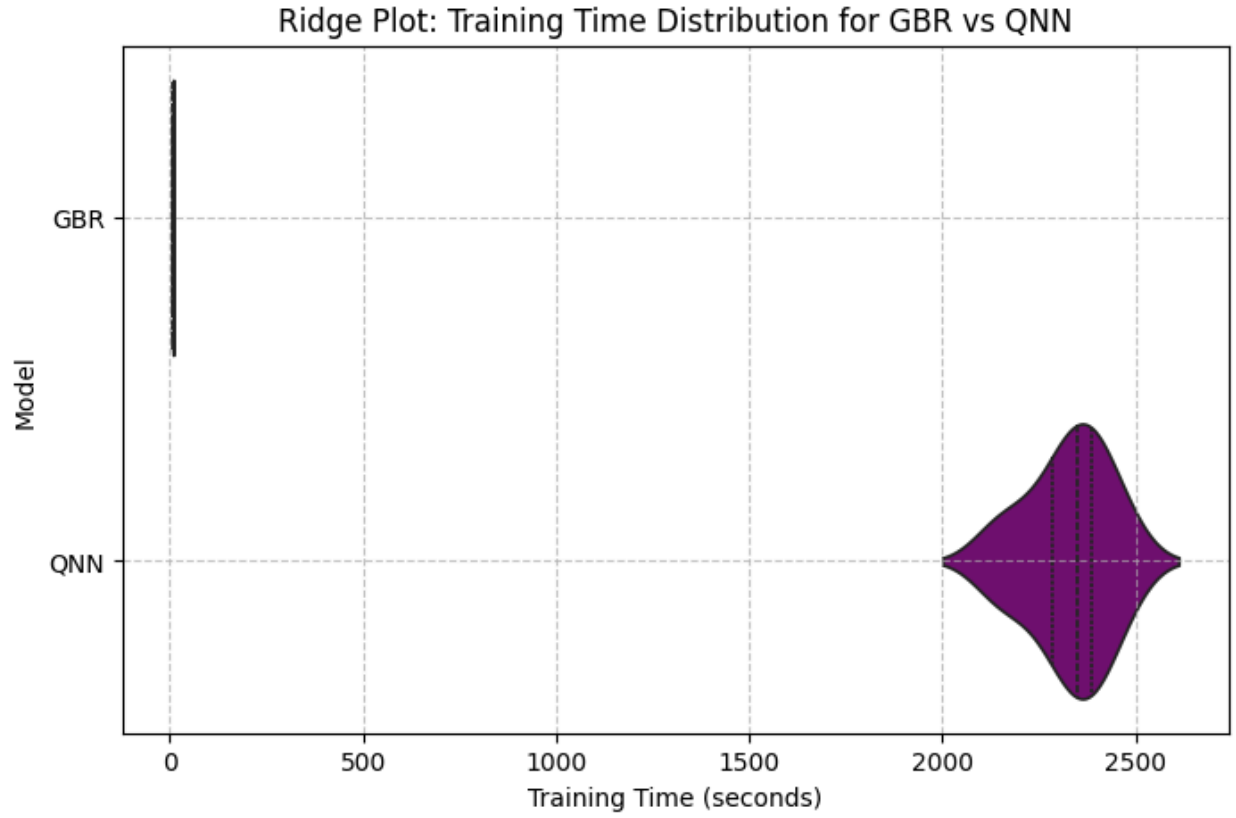
This graph shows how the training loss (MSE) changes over epochs for both the GBR and QNN models. Ideally, the loss should decrease as the model learns from the data. The rate of decrease and the final loss value provide insights into the model's training process and performance.

The GBR loss remains at 0 throughout the training process. This is unusual and suggests potential issues. A constant loss of 0 indicates that the GBR model is perfectly fitting the training data from the start. This might be a sign of overfitting or an error in the data or model setup.

The QNN loss decreases over epochs, showing that the model is learning from the data. The QNN loss converges to a low value, indicating that the model is able to fit the training data reasonably well.

The QNN model shows a typical learning curve with decreasing loss, suggesting that it is training properly. The constant zero loss for the GBR model is highly unusual and suggests a potential issue.

In conclusion, the graph suggests that the QNN model is learning properly, while the GBR model's constant zero loss is highly unusual and requires further investigation. The GBR model might be overfitting or have an error in the data or model setup.



The graph compares the training time distributions of two models (GBR and QNN) using a ridge plot. This visualization allows us to see the shape, spread, and central tendency of the training times for each model and compare them directly.

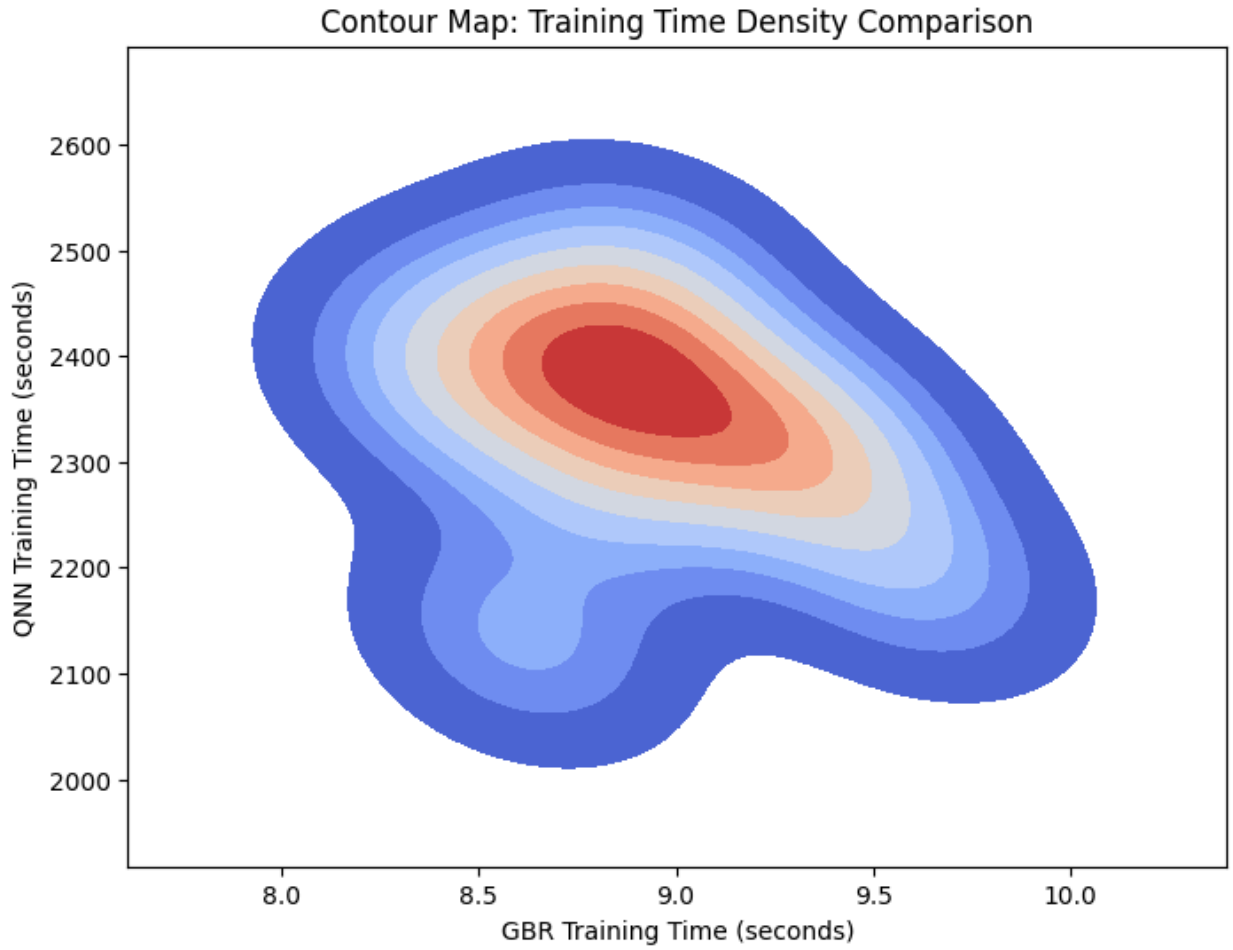
The GBR model shows a very narrow and concentrated distribution of training times near zero seconds. This indicates that the GBR model trains extremely quickly. The narrow shape suggests that the training time for GBR is very consistent across different runs or datasets.

The QNN model shows a much wider and shifted distribution of training times, ranging from approximately 2000 to 2600 seconds. This indicates that the QNN model takes significantly

longer to train compared to the GBR model. The wider shape suggests that the training time for QNN is more variable.

The GBR model is significantly more efficient in terms of training time compared to the QNN model. The QNN model has a much higher computational cost during training.

In conclusion, the ridge plot clearly shows a significant difference in training time between the GBR and QNN models. The GBR model trains extremely quickly, while the QNN model takes significantly longer and has more variability in training time. This suggests that the GBR model is more efficient in terms of training time.



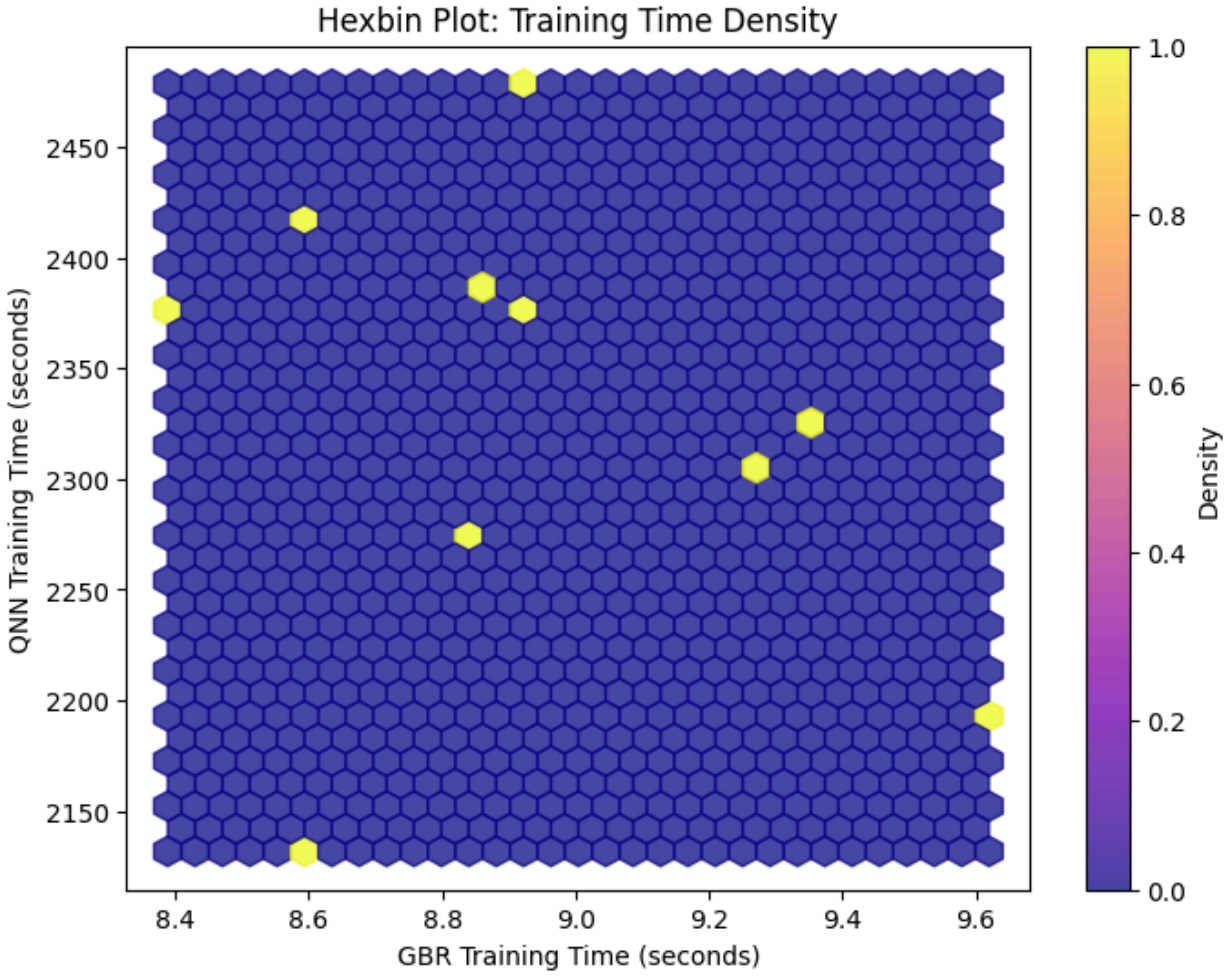
The graph aims to visualize the joint distribution of training times for GBR and QNN models. By showing the density of data points in a 2D space, it helps us understand the relationship between their training times and identify common patterns.

The GBR training times are concentrated in a very narrow range, approximately between 8 and 10 seconds. The QNN training times are concentrated in a much wider range, approximately between 2000 and 2600 seconds. The highest density area (red) shows that the most frequent combination is around 8.5 seconds for GBR and 2400 seconds for QNN. The contour plot

shows a negative correlation between GBR and QNN training times. When GBR training time is lower, QNN training time is higher, and vice versa. The graph highlights a significant difference in training times between the two models. GBR trains much faster than QNN.

The GBR model is significantly more efficient in terms of training time compared to the QNN model. The QNN model has a much higher computational cost during training. Training QNN models requires significantly more computational resources and time compared to GBR models.

In conclusion, the contour plot clearly shows a significant difference in training time between the GBR and QNN models. The GBR model trains extremely quickly, while the QNN model takes significantly longer and has more variability in training time. This suggests that the GBR model is more efficient in terms of training time. The negative correlation indicates that when GBR training time is low, QNN training time is high, and vice versa.



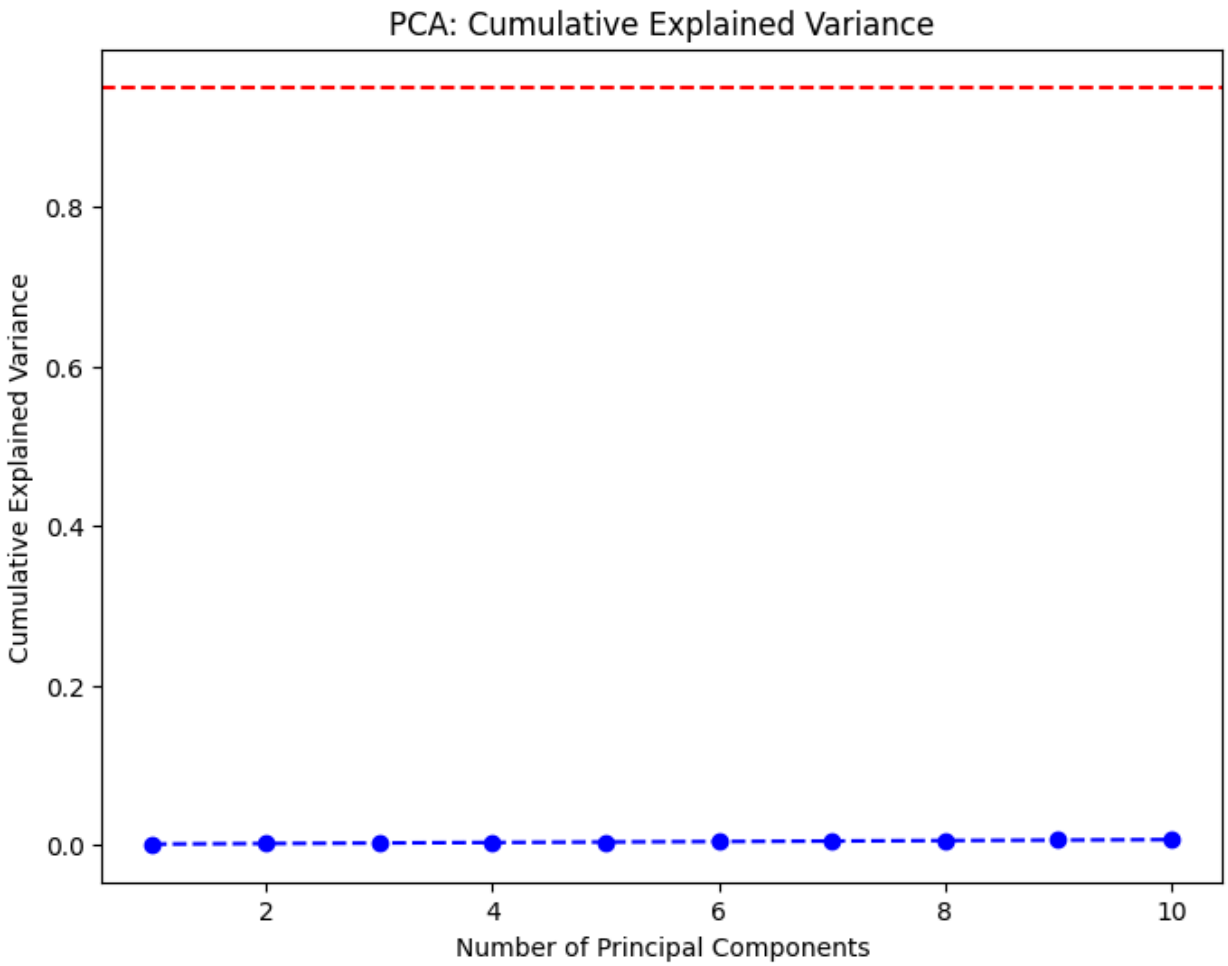
The graph aims to visualize the joint distribution of training times for GBR and QNN models. By showing the density of data points in hexagonal bins, it helps us understand the relationship between their training times and identify common patterns.

The GBR training times are concentrated in a very narrow range, approximately between 8.4 and 9.6 seconds. The QNN training times are concentrated in a much wider range, approximately between 2150 and 2475 seconds.

The graph shows that most of the bins have a low density (purple). This indicates that there are not many data points with specific combinations of GBR and QNN training times. There are a few scattered bins with higher density (yellow). These bins represent specific combinations of GBR and QNN training times that occurred more frequently. While not as clear as in a contour plot, the sparse distribution of high-density bins suggests a potential negative correlation between GBR and QNN training times. When GBR training time is lower, QNN training time tends to be higher, and vice versa. The graph highlights a significant difference in training times between the two models. GBR trains much faster than QNN.

The GBR model is significantly more efficient in terms of training time compared to the QNN model. The QNN model has a much higher computational cost during training. The few high-density bins might represent specific data points or scenarios where the training times deviated from the typical pattern.

In conclusion, the hexbin plot shows that the GBR model trains much faster than the QNN model. The sparse distribution of high-density bins suggests a potential negative correlation between their training times. The graph also indicates that there are not many data points with specific combinations of GBR and QNN training times, with a few potential outliers.

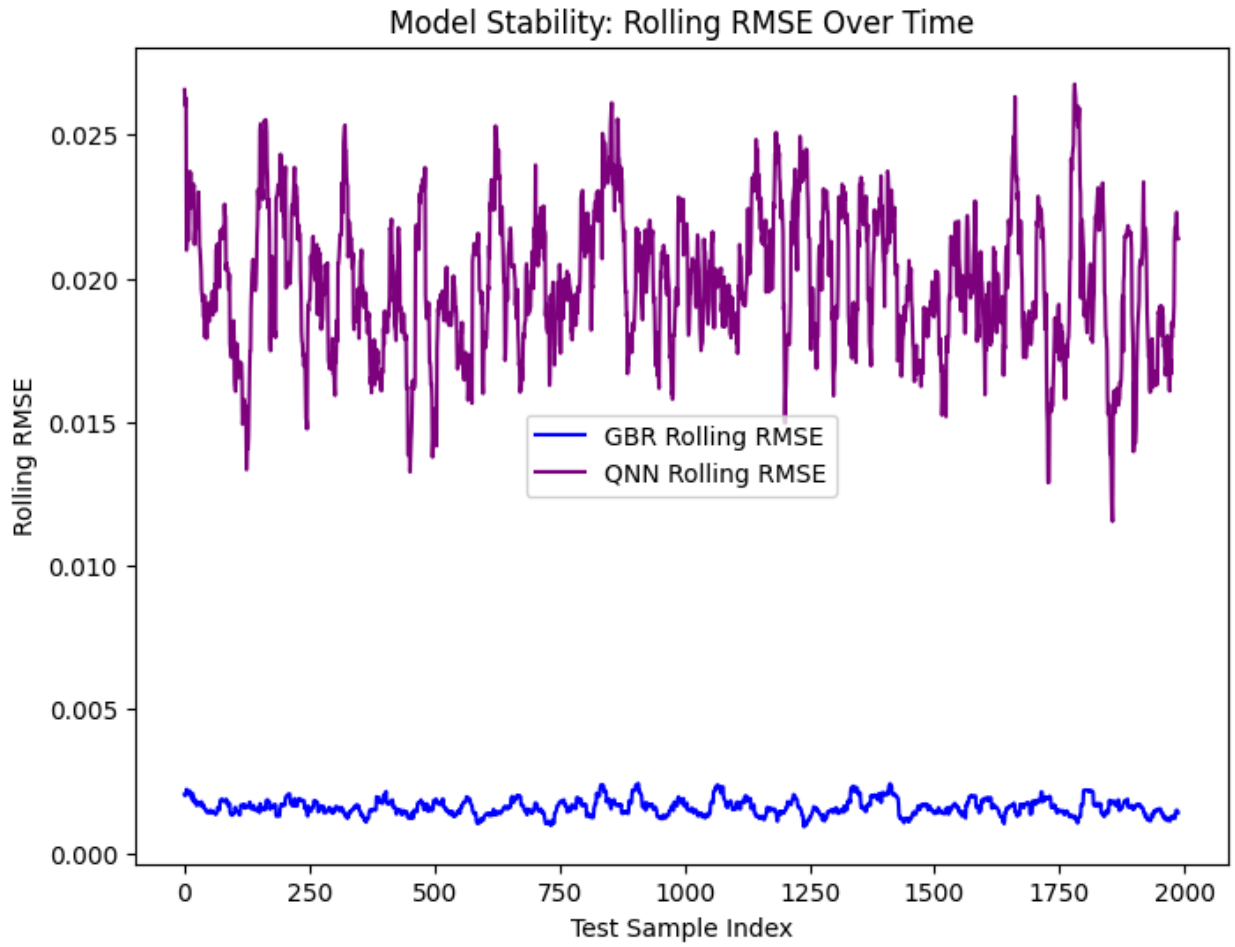


The graph aims to visualize how much of the original data's variance is captured as we include more principal components in the PCA. Ideally, we want to capture a large portion of the variance with a smaller number of components to achieve dimensionality reduction without losing too much information.

The blue line is very close to zero and shows minimal increase in cumulative explained variance as the number of principal components increases. This indicates that the principal components are not capturing much of the original data's variance. Even with 10 principal components, the cumulative explained variance remains close to zero. This suggests that the PCA is not effective in representing the data with fewer dimensions.

The red dashed line at 0.95 seems inconsistent with the data, as the cumulative explained variance never reaches that level. This suggests a potential misunderstanding or misinterpretation of the target variance. The graph indicates that PCA is not effective in reducing the dimensionality of this data. Alternative dimensionality reduction or feature selection methods might be more suitable for this data. The data and the application of PCA need to be re-evaluated to understand the reasons for the poor performance.

In conclusion, the graph shows that the cumulative explained variance remains very low even with an increasing number of principal components. This suggests that PCA is not effective in capturing the variance of the data and that alternative methods or a re-evaluation of the data and PCA application are needed.



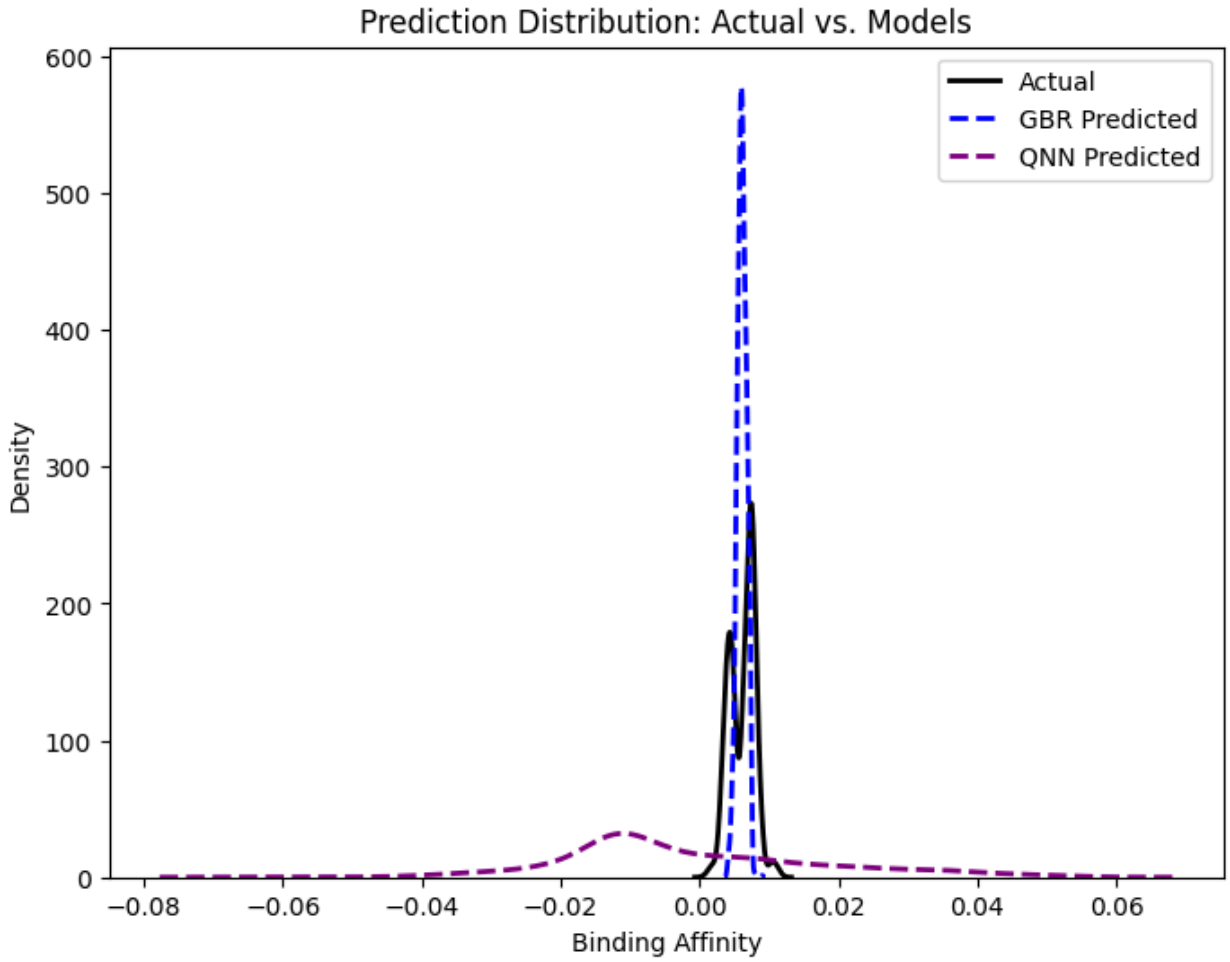
The graph aims to visualize the stability of the model's performance over time or across different parts of the test dataset. A stable model should have a consistent rolling RMSE, indicating that its performance doesn't fluctuate significantly.

The GBR rolling RMSE is consistently low and stable throughout the test sample index. This indicates that the GBR model's performance is stable and doesn't change much over time or across different subsets of the test data. The consistent low RMSE suggests that the GBR model is reliable and generalizes well to unseen data.

The QNN rolling RMSE is significantly higher than the GBR rolling RMSE and shows considerable fluctuations over the test sample index. This indicates that the QNN model's performance is unstable and varies significantly over time or across different subsets of the test data. The high and fluctuating RMSE suggests that the QNN model is less reliable and might not generalize well to unseen data.

The GBR model exhibits significantly better stability compared to the QNN model. The QNN model's instability raises concerns about its reliability and generalization ability.

In conclusion, the graph clearly shows that the GBR model exhibits significantly better stability compared to the QNN model. The GBR rolling RMSE is consistently low and stable, while the QNN rolling RMSE is high and fluctuates significantly. This suggests that the GBR model is more reliable and generalizes better to unseen data.



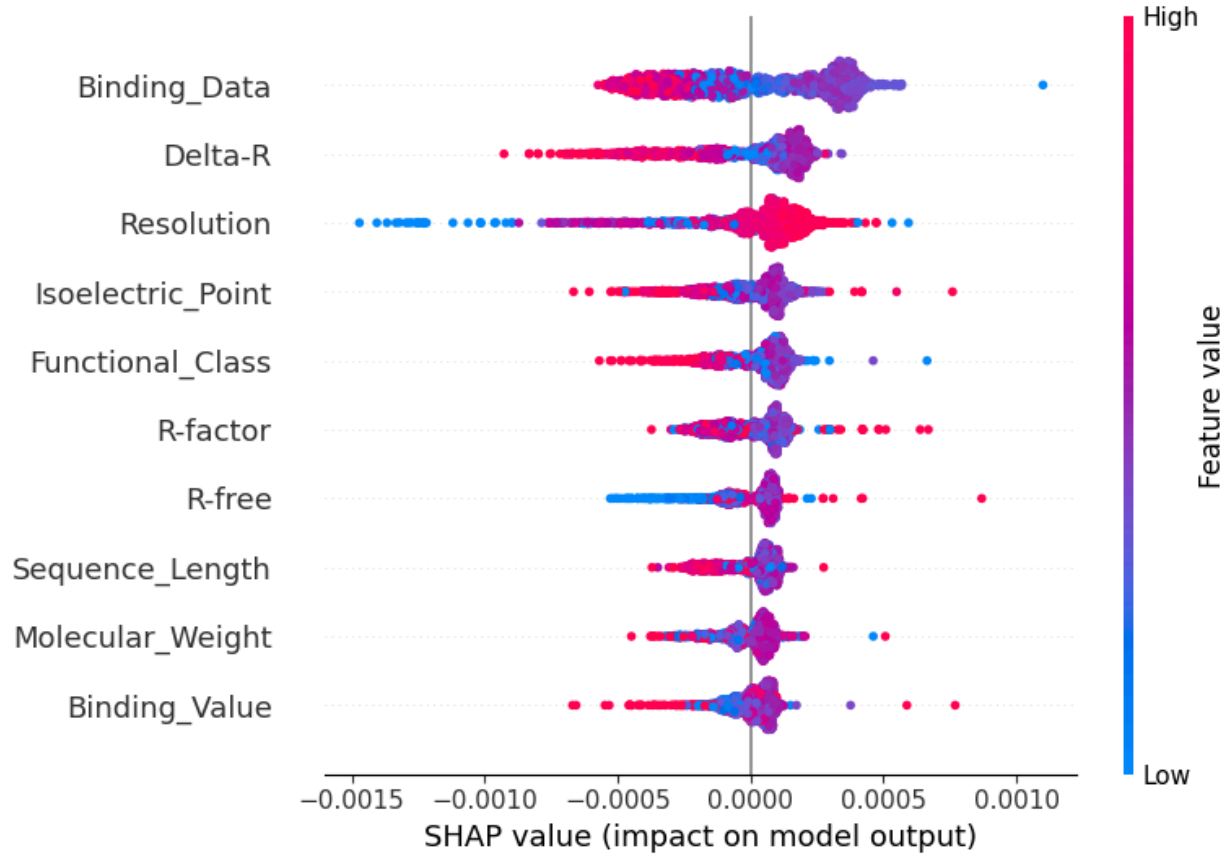
The graph compares the distribution of actual binding affinity values with the distributions of predicted values from two different models (GBR and QNN). Ideally, the predicted distributions should closely match the actual distribution, indicating accurate predictions.

The actual distribution has a sharp peak around zero binding affinity. This indicates that most of the actual values are concentrated near zero. The actual distribution has a low spread, suggesting that the actual values are relatively close to each other.

The GBR predicted distribution has a sharp peak that closely matches the actual distribution's peak. This indicates that the GBR model's predictions are well-aligned with the actual values. The GBR predicted distribution also has a low spread, similar to the actual distribution.

The QNN predicted distribution has a broader peak that is shifted away from zero, towards the negative side. This indicates that the QNN model's predictions are less accurate and have a systematic bias towards underestimating the values. The QNN predicted distribution has a higher spread compared to the actual and GBR predicted distributions.

In conclusion, the graph shows that the GBR model's predictions closely match the actual distribution, indicating high accuracy. The QNN model's predictions deviate significantly from the actual distribution, indicating lower accuracy and a systematic bias. This suggests that the GBR model is a better choice for predicting binding affinity values.

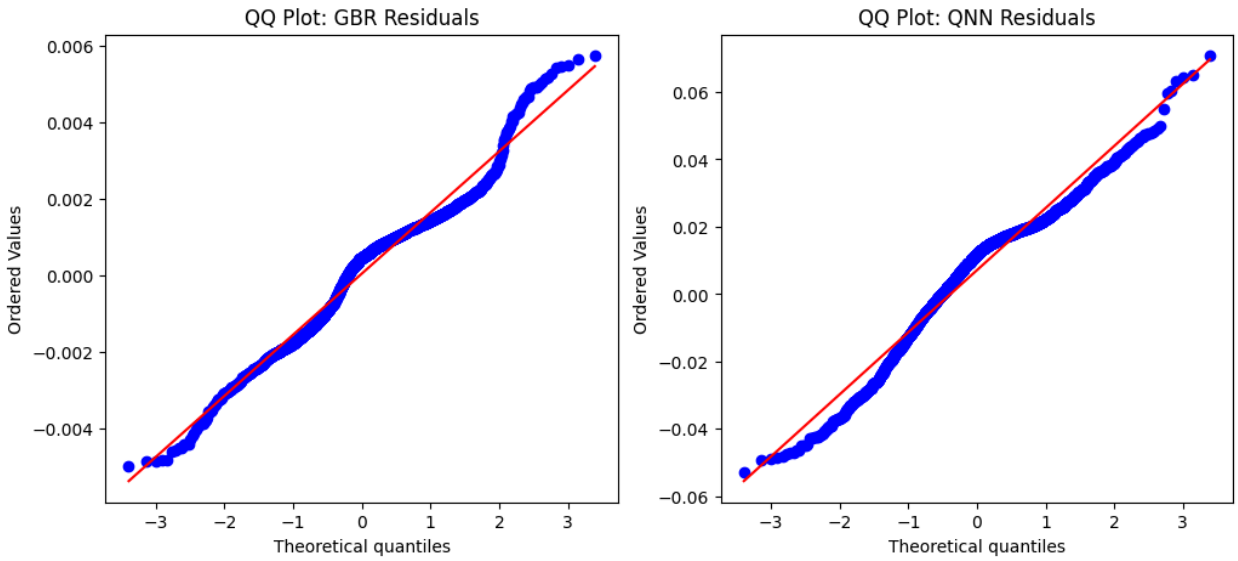


SHAP values are based on game theory and provide a way to understand how each feature contributes to the model's prediction for a specific instance. The summary plot aggregates these instance-level explanations to show the overall importance and impact of each feature.

The features at the top of the plot have the most significant impact on the model's output. In this case, "Binding_Data" is the most important feature, followed by "Delta-R" and "Resolution". High values of "Binding_Data" have a positive SHAP value, meaning they push the prediction higher.

Low values have a negative SHAP value, pushing the prediction lower. High values of "Delta-R" have a positive SHAP value, pushing the prediction higher. Low values have a negative SHAP value, pushing the prediction lower. Low values of "Resolution" have a positive SHAP value, pushing the prediction higher. High values have a negative SHAP value, pushing the prediction lower. The color of the dots shows the distribution of feature values. For example, "Binding_Data" has a mix of high and low values, while "Resolution" has more low values. The width of the distribution of SHAP values for a feature indicates the range of its impact on the model's output. "Binding_Data" has a wider range of impact compared to "R-free", for example. The plot can reveal relationships between features and the model's output. For example, "Binding_Data" and "Delta-R" have a similar pattern of impact, suggesting they might be correlated.

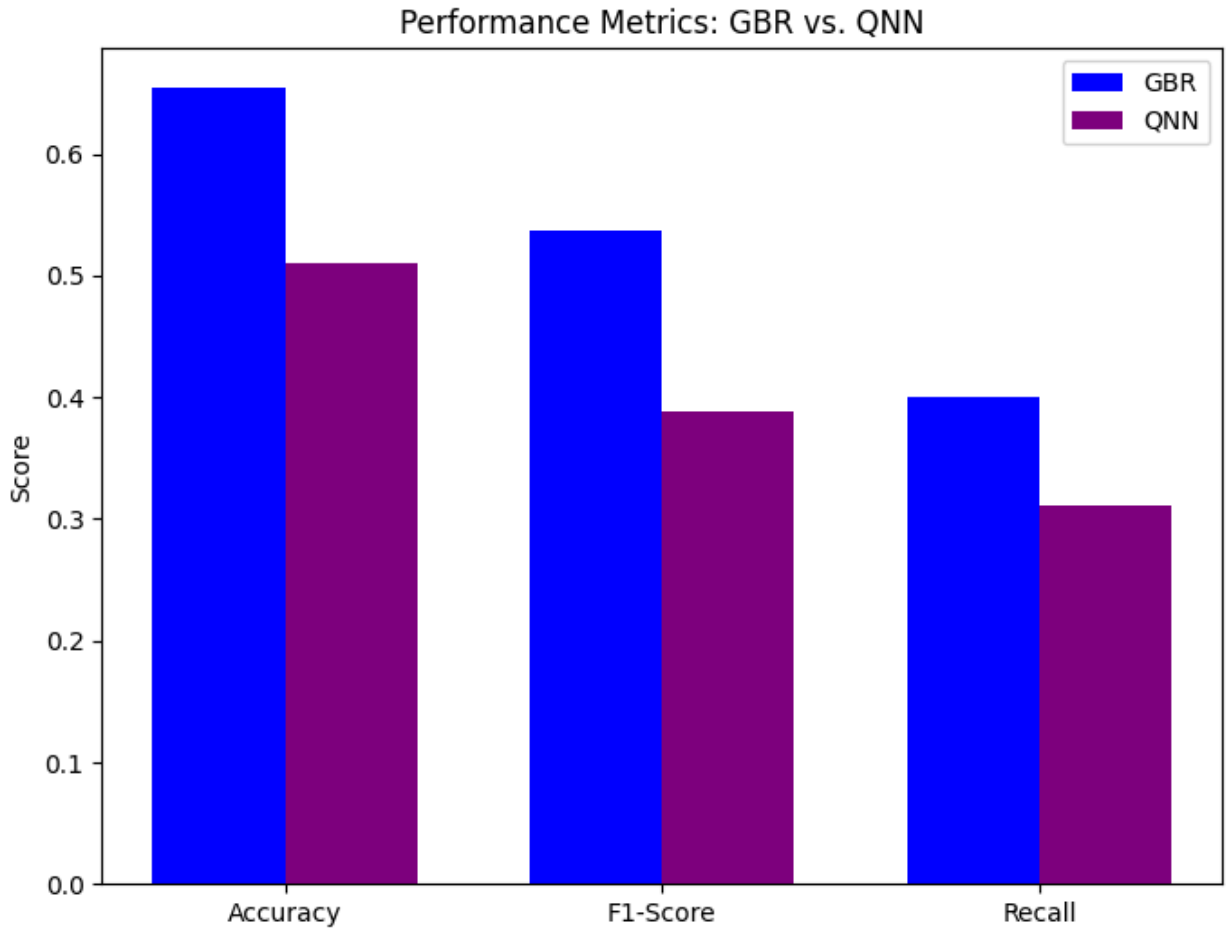
In conclusion, the SHAP summary plot provides valuable insights into the importance and impact of features in the CatBoost model. It shows that "Binding_Data", "Delta-R", and "Resolution" are the most important features, and it reveals the direction and range of their impact on the model's predictions. This information can be used to understand the model's behavior, identify important features, and potentially improve the model's performance.



QQ plots are used to assess whether a dataset (in this case, the residuals) follows a particular distribution (in this case, a normal distribution). If the residuals are normally distributed, the blue dots should fall closely along the red line. Deviations from the line indicate departures from normality.

The blue dots in the GBR plot show some deviations from the red line, particularly at the tails. This suggests that the GBR residuals are not perfectly normally distributed. The pattern of the blue dots shows a slight S-shape, which is a common indication of non-normality. This suggests that the tails of the residual distribution are heavier than a normal distribution.

Closer to the Red Line: The blue dots in the QNN plot are generally closer to the red line compared to the GBR plot. This suggests that the QNN residuals are closer to being normally distributed than the GBR residuals. However, there are still some deviations from the red line, particularly at the tails, indicating that the QNN residuals are not perfectly normally distributed either.n conclusion, the QQ plots reveal that the QNN residuals are closer to being normally distributed compared to the GBR residuals. The GBR residuals show a more significant deviation from normality, particularly the S-shape pattern. This suggests that the QNN model might be more appropriate for statistical methods that assume normality, while the GBR model might require further investigation or model improvement to address the non-normality of residuals.

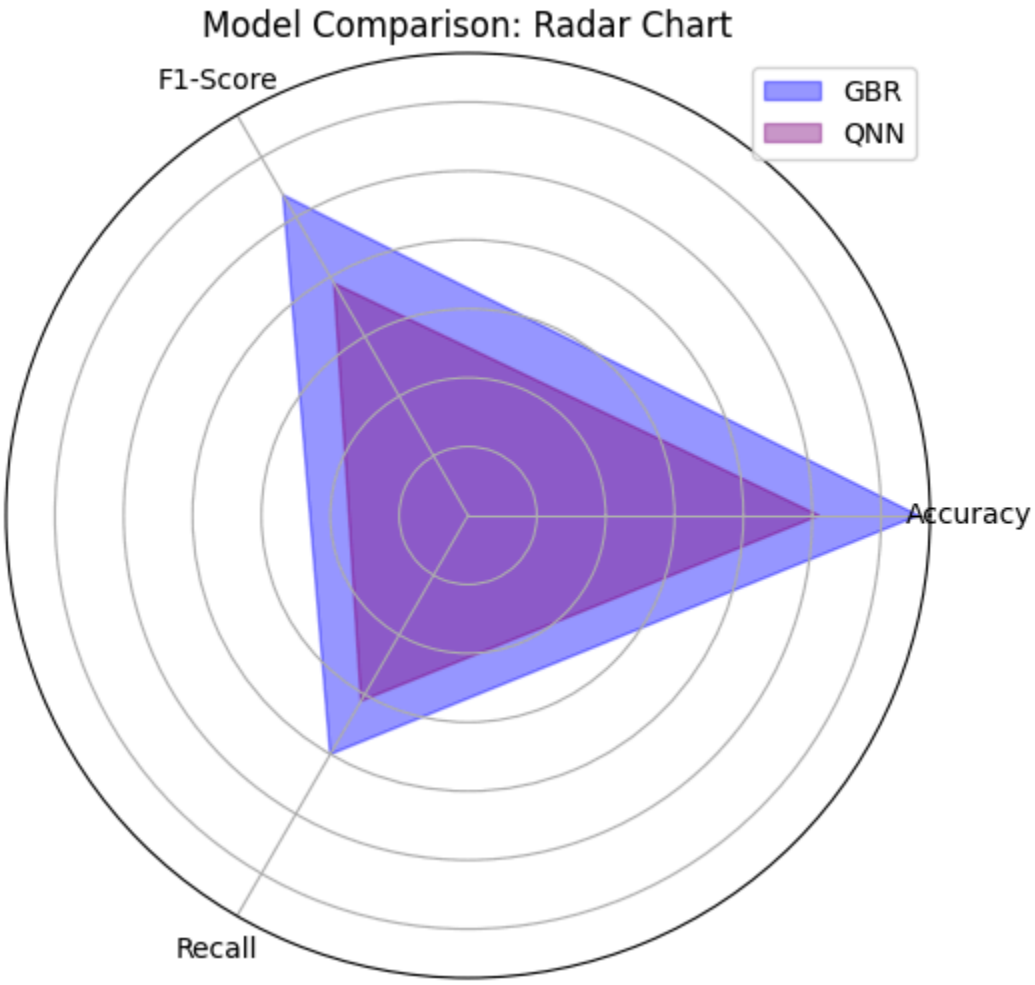


The graph compares the performance of two models (GBR and QNN) based on three common classification metrics. The goal is to determine which model performs better based on these metrics.

The GBR model has a higher accuracy compared to the QNN model. This indicates that the GBR model correctly predicts a larger proportion of instances. The GBR model has a higher F1-score compared to the QNN model. This suggests that the GBR model has a better balance

between precision and recall. The GBR model has a higher recall compared to the QNN model. This indicates that the GBR model is better at identifying actual positive instances.

In conclusion, the graph clearly shows that the GBR model outperforms the QNN model in all three metrics: accuracy, F1-score, and recall. This indicates that the GBR model is generally better at classifying instances for this particular task.

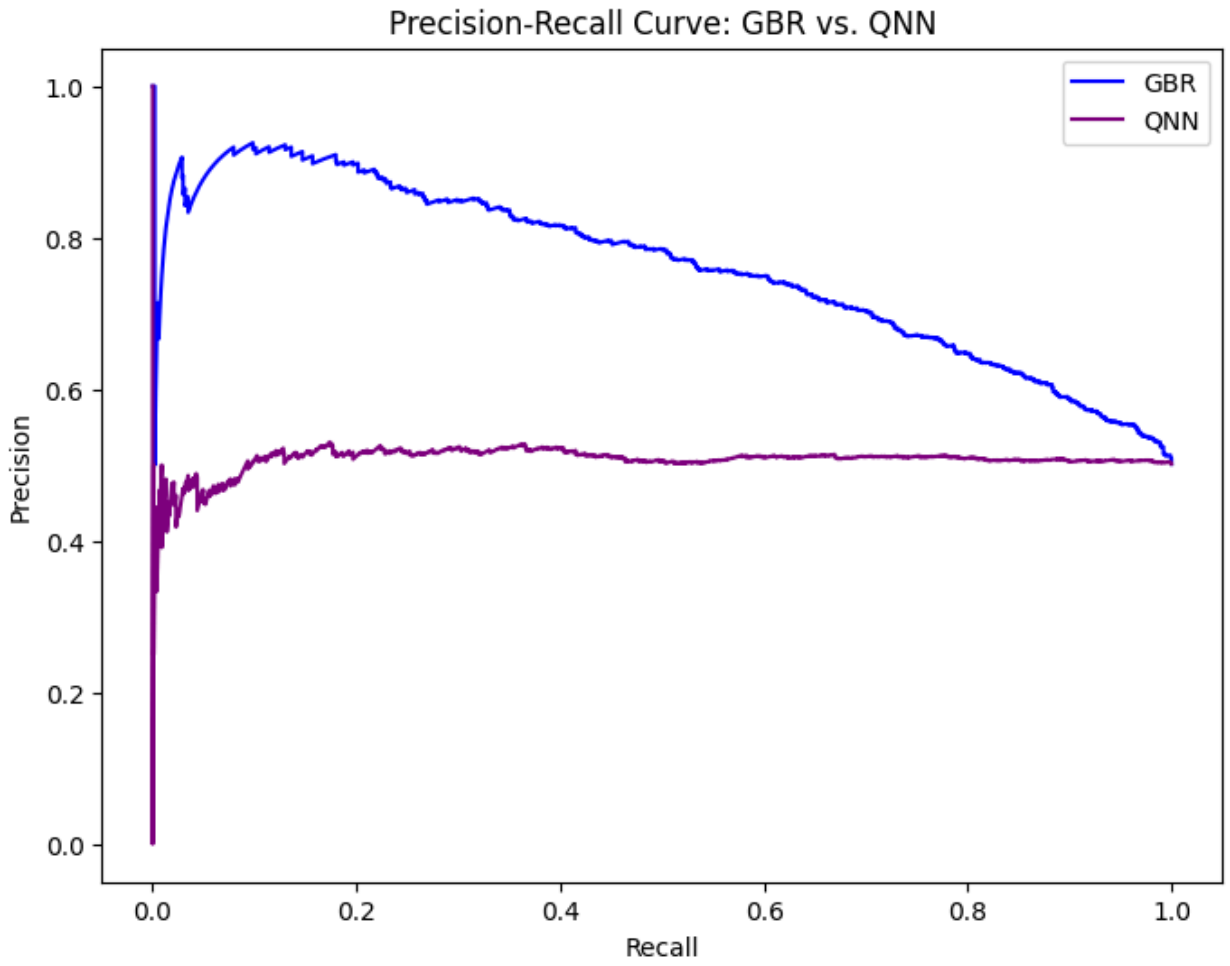


The radar chart provides a visual comparison of the performance of two models (GBR and QNN) across multiple metrics. The shape and size of the polygons allow for a quick assessment of each model's strengths and weaknesses.

The blue polygon (GBR) encloses a larger area compared to the purple polygon (QNN). This indicates that the GBR model has better overall performance across all three metrics. The GBR polygon extends further towards the outer circles on all three axes, indicating higher scores for Accuracy, F1-Score, and Recall.

The purple polygon (QNN) encloses a smaller area compared to the blue polygon (GBR). This indicates that the QNN model has lower overall performance across all three metrics. The QNN

polygon is closer to the center on all three axes, indicating lower scores for Accuracy, F1-Score, and Recall. In conclusion, the radar chart clearly shows that the GBR model outperforms the QNN model in terms of Accuracy, F1-Score, and Recall. The larger area enclosed by the GBR polygon indicates better overall performance. This suggests that the GBR model is a more suitable choice for this classification task.

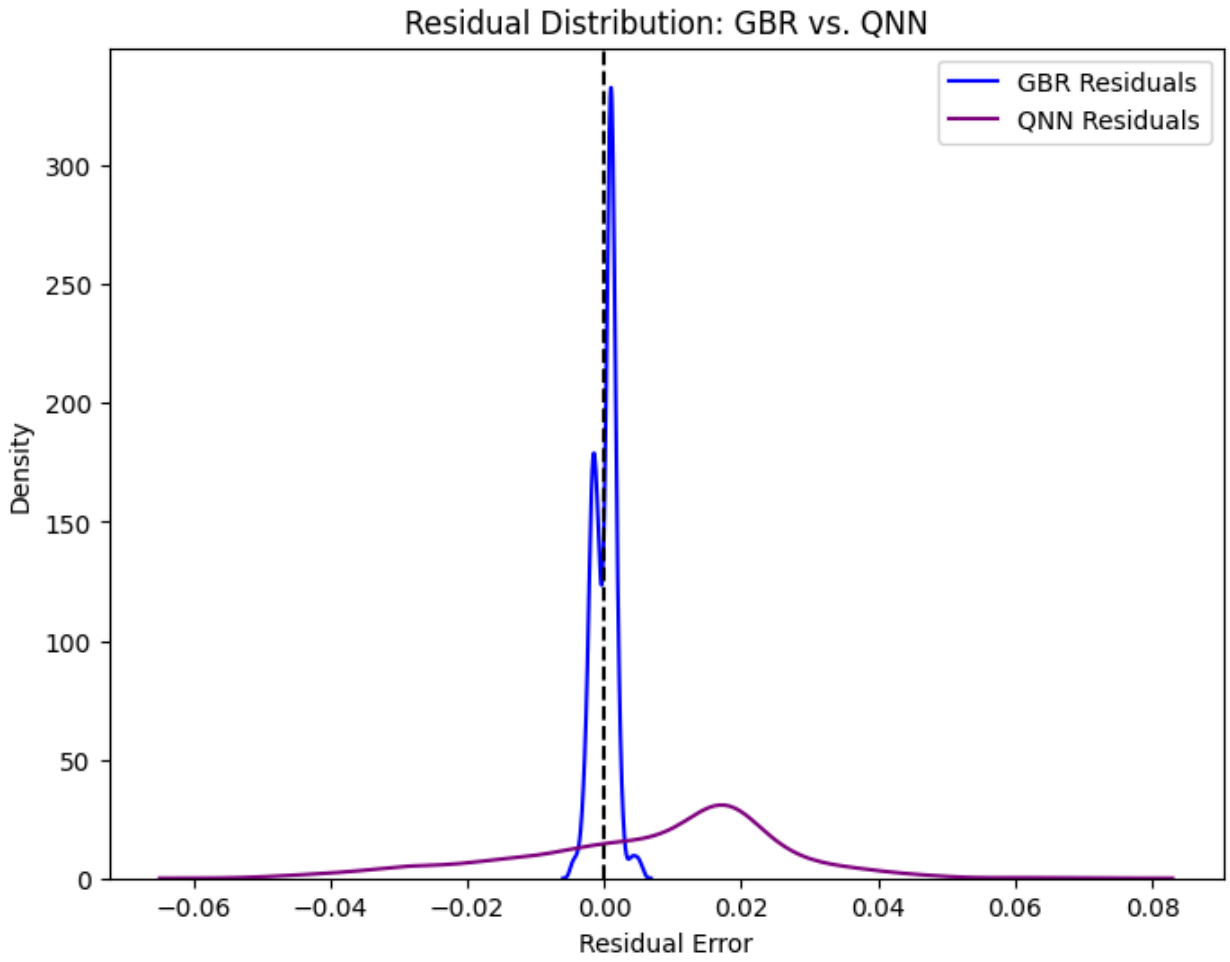


The Precision-Recall (PR) curve shows the trade-off between precision and recall for different threshold values. A high area under the curve (AUC) represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. A high AUC indicates that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall).

The GBR curve (blue line) is consistently higher than the QNN curve across the entire recall range. This means that for any given recall value, the GBR model achieves higher precision. Since the GBR curve is consistently above the QNN curve, it implies that the GBR model has a

higher area under the PR curve (AUC-PR). The GBR model demonstrates better performance in terms of the precision-recall trade-off.

The QNN curve (purple line) is consistently lower than the GBR curve, indicating lower precision for any given recall value. The QNN model has a lower implied AUC-PR compared to the GBR model. The QNN model exhibits poorer performance in terms of the precision-recall trade-off. In conclusion, the Precision-Recall curve clearly shows that the GBR model outperforms the QNN model. The GBR model achieves higher precision across all recall values, indicating a better balance between precision and recall. This suggests that the GBR model is a more suitable choice for this classification task.

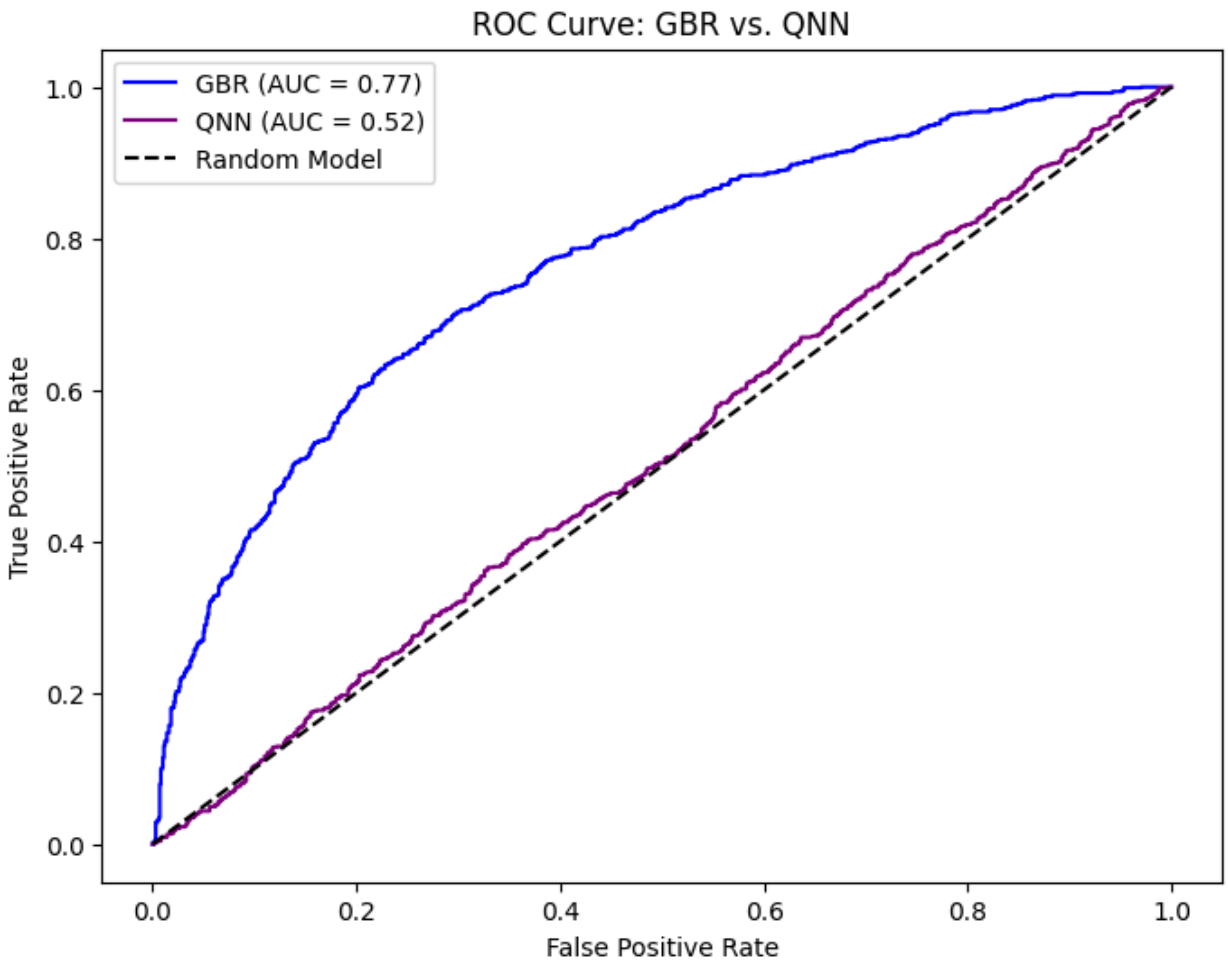


This graph compares the residual distributions of two different models (GBR and QNN). Ideally, the residuals should be centered around zero, indicating that the model is unbiased and makes accurate predictions on average. The shape and spread of the distributions provide insights into the model's performance.

The GBR residuals have a very narrow and high peak centered close to zero. This indicates that the GBR model's predictions are highly concentrated around the actual values, resulting in small residuals. The GBR residual distribution has a very low spread, meaning that most of the residuals are close to zero.

The QNN residuals have a broader peak that is shifted away from zero, towards the positive side. This indicates that the QNN model's predictions have a larger spread and are systematically biased, overestimating the values. The QNN residual distribution has a higher spread compared to the GBR residuals, meaning that the residuals vary more widely.

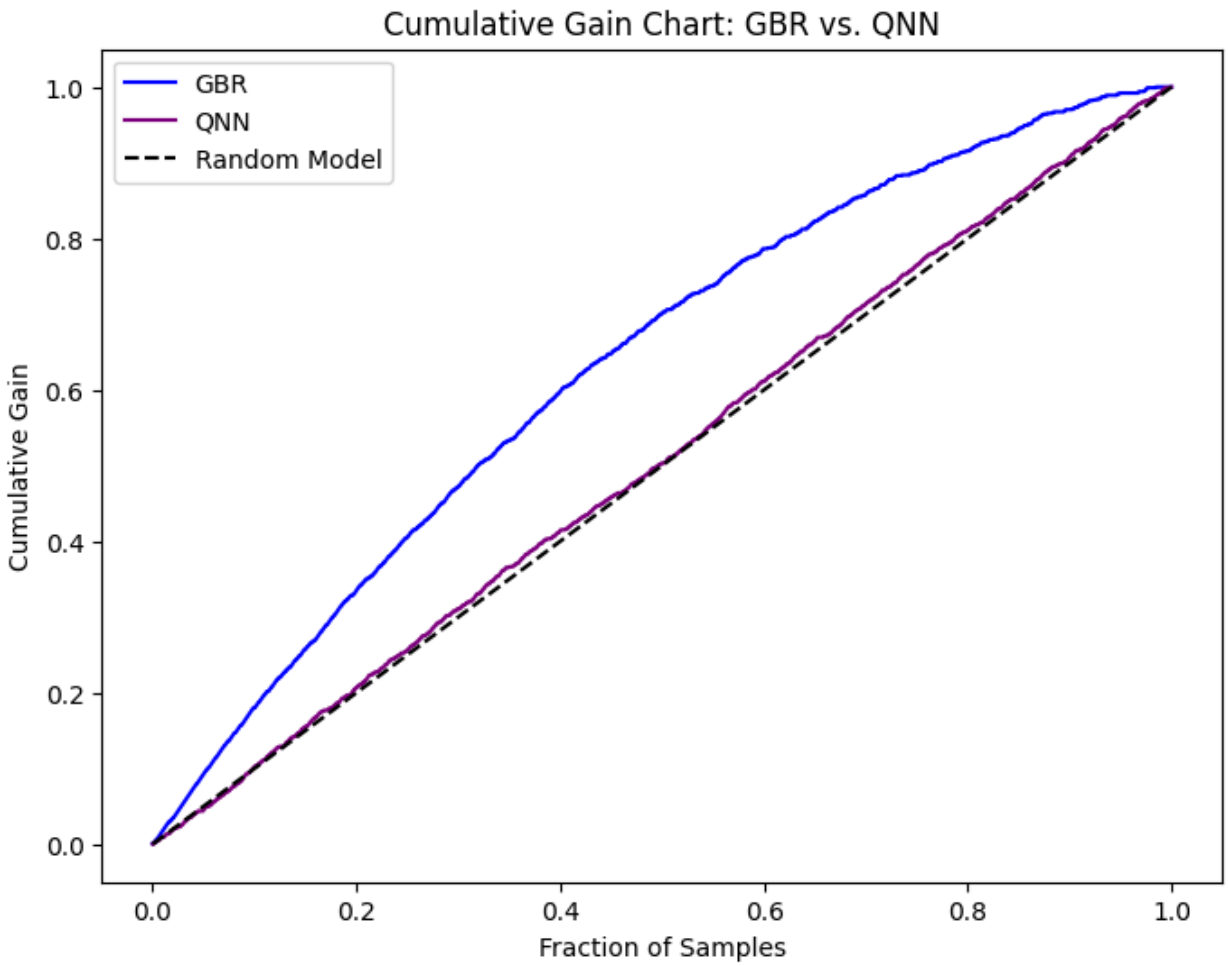
The GBR model significantly outperforms the QNN model in terms of prediction accuracy. The GBR residuals are concentrated around zero and have a low spread, indicating high accuracy and low bias. The QNN model's residuals are broader and shifted away from zero, indicating lower accuracy and a systematic bias. In conclusion, the "Residual Distribution: GBR vs. QNN" graph clearly shows that the GBR model has significantly better performance than the QNN model. The GBR residuals are concentrated around zero and have a low spread, indicating high accuracy and low bias. The QNN model's residuals are broader and shifted away from zero, indicating lower accuracy and a systematic bias.



The ROC curve shows the trade-off between the true positive rate (TPR) and the false positive rate (FPR) for different threshold values. A good model aims to maximize TPR while minimizing FPR, resulting in a curve that is closer to the top-left corner.

The GBR model has a higher AUC (0.77) compared to the QNN model. The GBR curve (blue line) is closer to the top-left corner, indicating better performance in terms of distinguishing between positive and negative classes. The GBR model demonstrates better overall performance based on the ROC curve and AUC.

The QNN model has a lower AUC (0.52), which is very close to the AUC of a random model (0.5). The QNN curve (purple line) is closer to the random model line, indicating poor performance. The QNN model exhibits poorer performance based on the ROC curve and AUC. In conclusion, the ROC curve clearly shows that the GBR model outperforms the QNN model. The GBR model has a higher AUC and a curve that is closer to the top-left corner, indicating better performance. The QNN model's performance is close to that of a random model, suggesting that it has little to no predictive power.

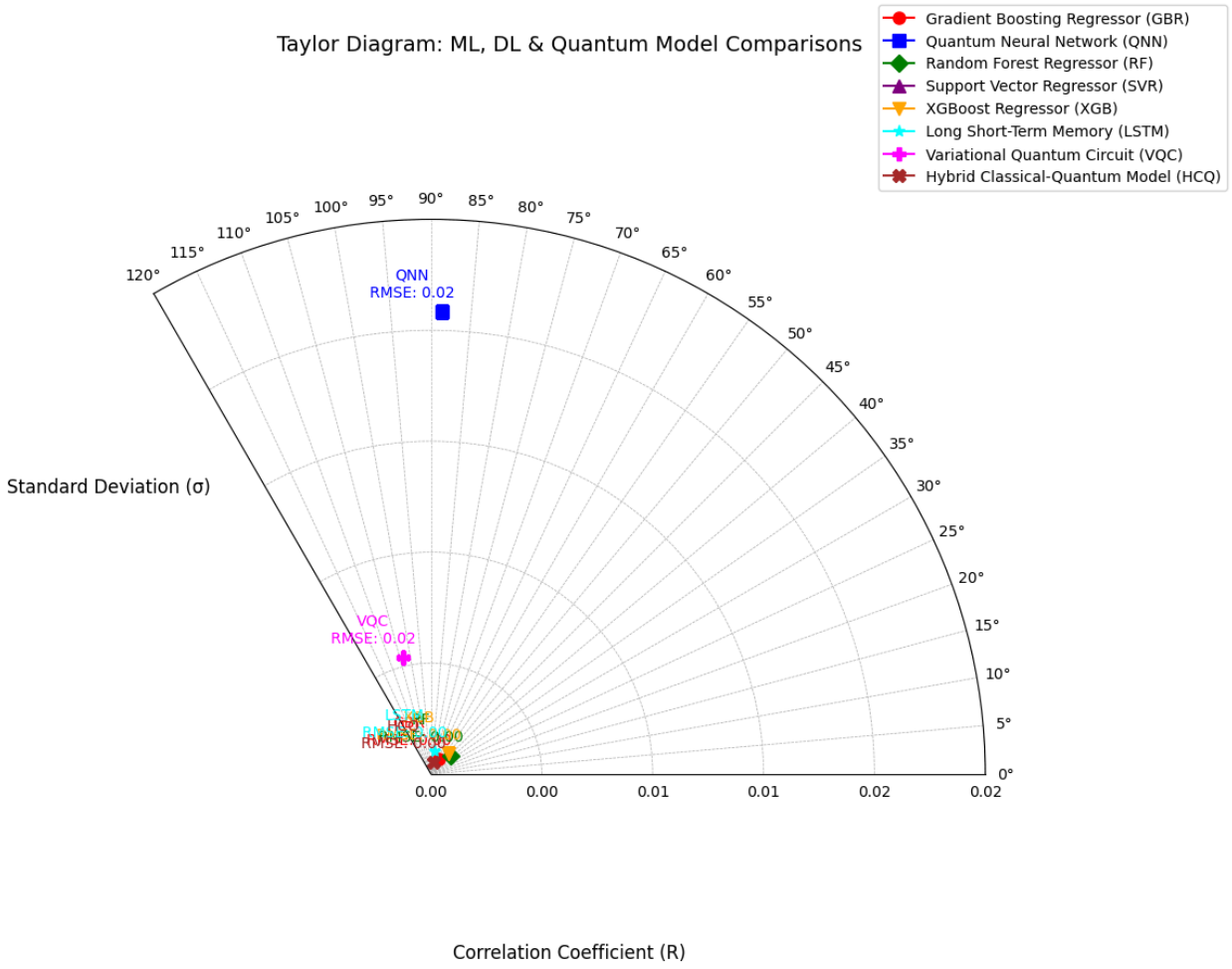


The cumulative gain chart is used to evaluate the performance of a classification model, particularly in scenarios where the goal is to identify a small subset of positive samples within a larger dataset. A good model will show a steeper curve in the earlier fractions of samples, indicating that it is effectively capturing a larger proportion of the positives with fewer samples.

The GBR curve (blue line) is steeper than the random model line and the QNN curve in the earlier fractions of samples. This indicates that the GBR model is effectively capturing a larger proportion of the positives with fewer samples. The GBR model shows a higher cumulative gain compared to both the random model and the QNN model for the same fraction of samples. The GBR model demonstrates better overall performance in terms of cumulative gain.

The QNN curve (purple line) is closer to the random model line, indicating that its performance is closer to random guessing. The QNN model shows a lower cumulative gain compared to the GBR model for the same fraction of samples. The QNN model exhibits poorer performance in terms of cumulative gain. In conclusion, the cumulative gain chart clearly shows that the GBR model outperforms the QNN model. The GBR model achieves a higher cumulative gain in the earlier fractions of samples, indicating that it is more effective at capturing positive samples. The

QNN model's performance is closer to that of a random model, suggesting that it has limited predictive power.



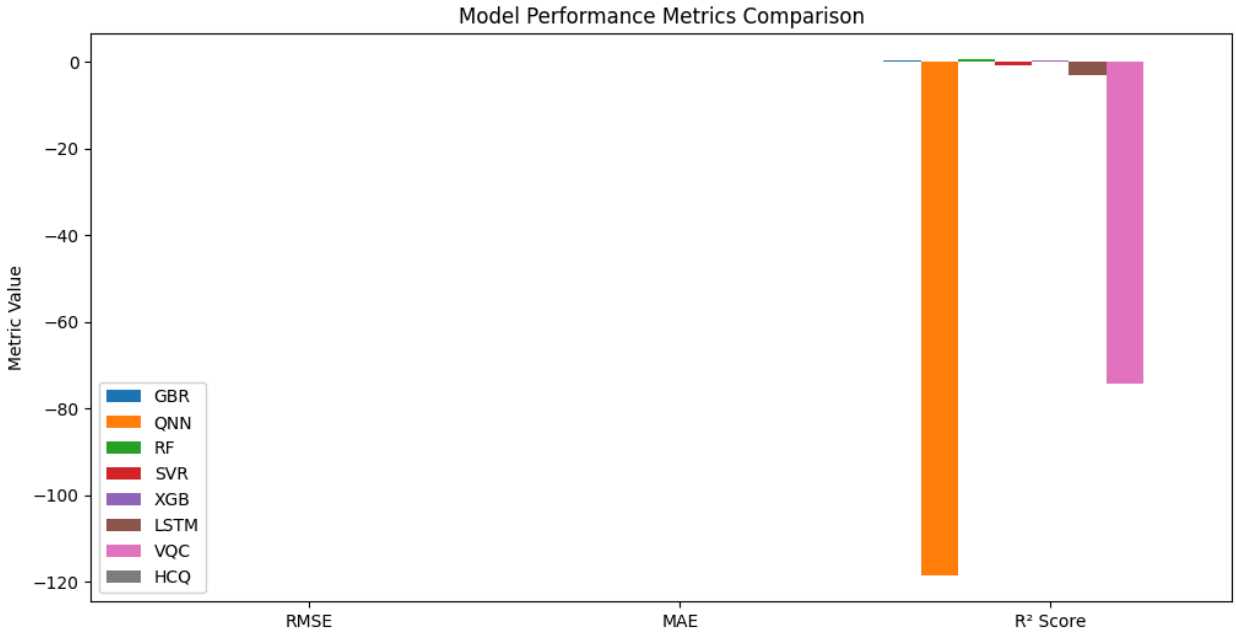
The Taylor Diagram combines three statistical measures (correlation, RMSE, and standard deviation) into a single plot, allowing for a comprehensive comparison of model performance. The closer a point is to the reference point (1, 0), the better the model's performance.

Most models (GBR, RF, SVR, XGB, LSTM, and HCQ) are clustered very close to the reference point (1, 0). This indicates that these models have high correlation, low RMSE, and similar standard deviation to the actual data, signifying good performance.The Quantum Neural Network (QNN) and Variational Quantum Circuit (VQC) models are located further away from the reference point. This indicates that these models have lower correlation, higher RMSE, or different standard deviations compared to the other models, signifying poorer performance.

GBR, RF, SVR, XGB, LSTM, HCQ these models show very good performance, with high correlation and low RMSE. There is little to distinguish between them in this diagram, suggesting they perform similarly well. The QNN model has a significantly higher RMSE (0.02) compared to the other models. It also has a lower correlation (not clearly visible but implied by

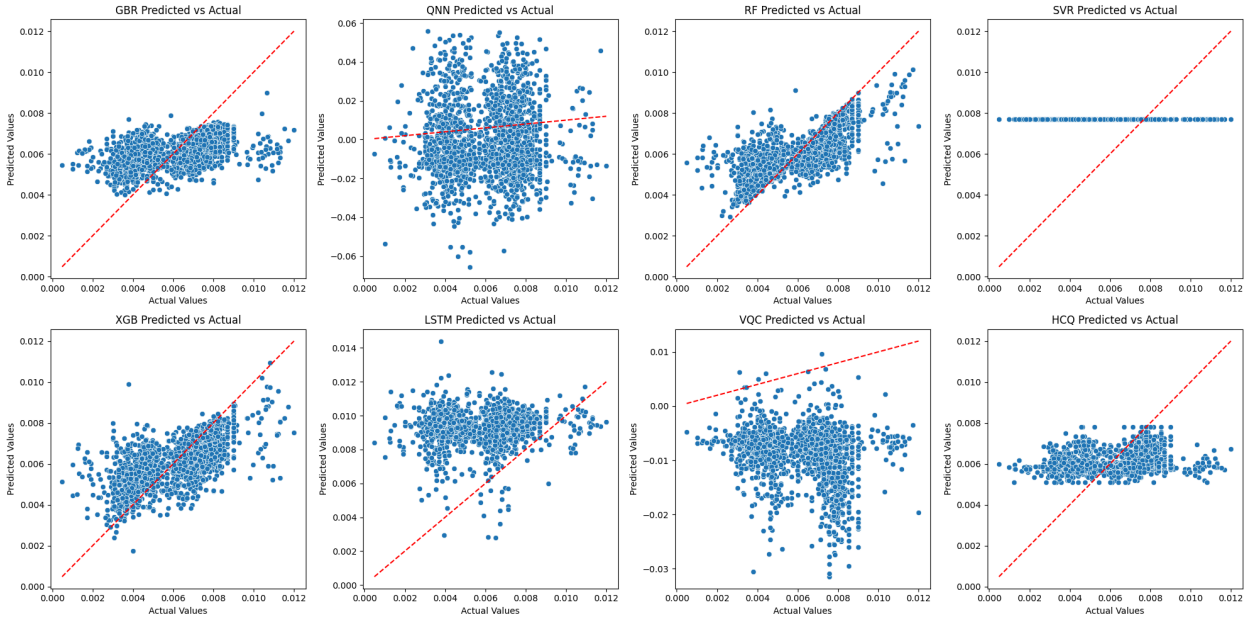
its position). The VQC model also has a higher RMSE (0.02) compared to the other models, though it is closer to the reference point than QNN.

The classical machine learning models (GBR, RF, SVR, XGB, LSTM) and the hybrid classical-quantum model (HCQ) demonstrate superior performance compared to the pure quantum models (QNN and VQC). In conclusion, the Taylor Diagram reveals that the classical machine learning models and the hybrid classical-quantum model demonstrate superior performance compared to the pure quantum models. The QNN and VQC models show significantly higher RMSE and lower correlation, indicating poorer performance. This suggests that the classical models are more effective for this particular task, and further investigation is needed to understand and address the issues with the quantum models.



The graph compares the performance of multiple models based on three common regression metrics. The goal is to determine which model performs better based on these metrics.

GBR, RF, SVR, XGB, LSTM, HCQ have RMSE of 0. So, these models have an RMSE of 0. This indicates perfect predictions with no errors. QNN and VQC have RMSE of -120, these models have an extremely large negative RMSE. This is highly unusual and suggests a serious problem with the model's predictions or the calculation of RMSE. In conclusion, the graph presents highly unusual metric values, particularly for the QNN and VQC models. The extremely large negative values suggest a severe problem with the models' predictions or the calculation of the metrics. The perfect scores for the other models are also suspicious and may indicate an error or a trivial prediction scenario. Further investigation is needed to understand and address these issues.



The fundamental logic behind these graphs is to assess the performance of different models in predicting values. If a model were perfect, all the blue dots would fall exactly on the red dashed line. The degree to which the dots deviate from this line indicates the model's prediction error or inaccuracy.

GBR (Gradient Boosting Regressor): The blue dots show moderate scatter around the red dashed line, indicating reasonable but not perfect predictions. There's a tendency for the model to slightly underestimate higher actual values.

QNN (Quantum Neural Network): The blue dots are significantly scattered, showing poor agreement between predicted and actual values. The predicted values are on a different scale, with many negative values, which is unusual for a binding affinity prediction task.

RF (Random Forest): Similar to GBR, the RF model shows moderate scatter, indicating reasonable but not perfect predictions. Again, there's a slight tendency to underestimate higher actual values.

SVR (Support Vector Regressor): The model shows a very poor fit, with predicted values clustered around a single point and not reflecting the variations in actual values.

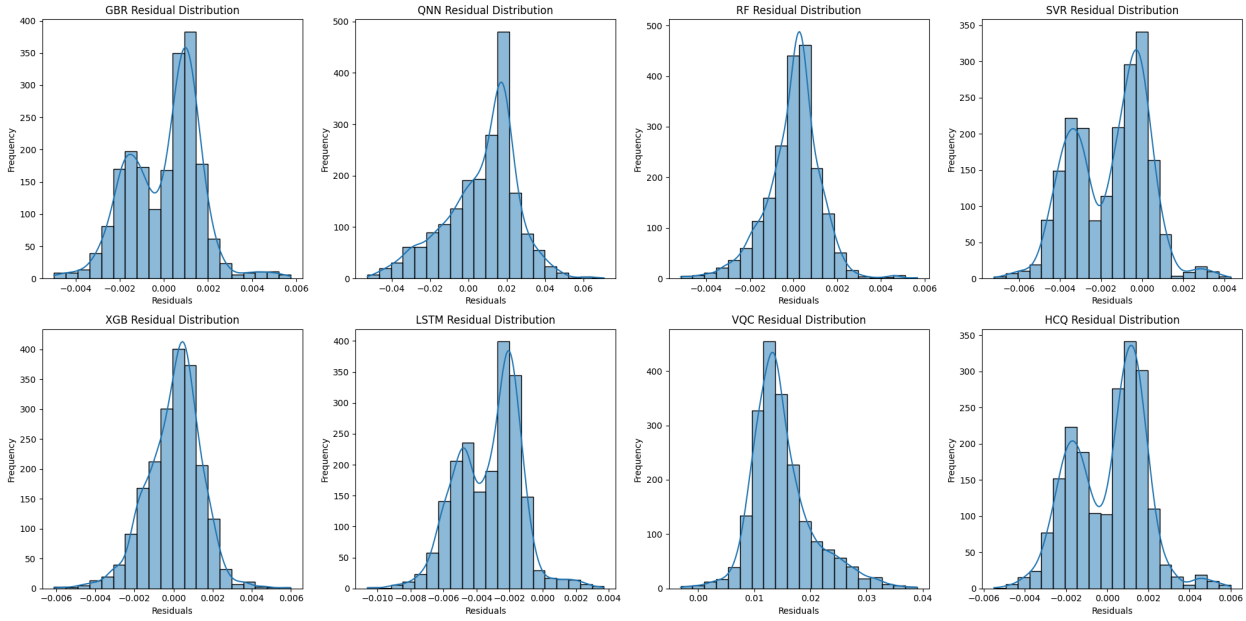
XGB (XGBoost): The XGB model shows moderate scatter, indicating reasonable but not perfect predictions. Similar to GBR and RF, there's a slight underestimation of higher actual values.

LSTM (Long Short-Term Memory): The LSTM model shows moderate scatter, indicating reasonable but not perfect predictions. Again, there's a slight underestimation of higher actual values.

VQC (Variational Quantum Circuit): The VQC model shows significant scatter, indicating poor agreement between predicted and actual values. The predicted values are on a different scale, with many negative values, similar to QNN.

HCQ (Hybrid Classical-Quantum): The HCQ model shows moderate scatter, indicating reasonable but not perfect predictions.Similar to GBR, RF, XGB, and LSTM, there's a slight underestimation of higher actual values.

In conclusion, the GBR, RF, XGB, LSTM, and HCQ models demonstrate reasonable performance, while the QNN, VQC, and SVR models show poor performance. The consistent underestimation bias in the better-performing models should be investigated and addressed.



The graph aims to visualize the distribution of prediction errors. Ideally, the errors should be centered around zero, indicating that the model is unbiased and makes accurate predictions on average. The shape and spread of the distribution provide insights into the model's performance.

GBR (Gradient Boosting Regressor): The residuals show a bimodal distribution with two peaks, suggesting the model might be struggling with certain subsets of the data. The distribution is slightly skewed to the right, indicating a tendency to underestimate.

QNN (Quantum Neural Network): The residuals show a unimodal distribution with a single peak. The distribution is slightly skewed to the right, indicating a tendency to underestimate. The residuals have a wider spread compared to GBR, suggesting higher variability in errors.

RF (Random Forest): The residuals show a unimodal distribution with a single peak. The distribution is slightly skewed to the right, indicating a tendency to underestimate. The residuals have a similar spread to QNN.

SVR (Support Vector Regressor): The residuals show a bimodal distribution with two distinct peaks, indicating potential issues with the model's fit. The residuals have a high spread, suggesting significant variability in errors.

XGB (XGBoost): The residuals show a unimodal distribution with a single peak. The distribution is slightly skewed to the right, indicating a tendency to underestimate. The residuals have a similar spread to QNN and RF.

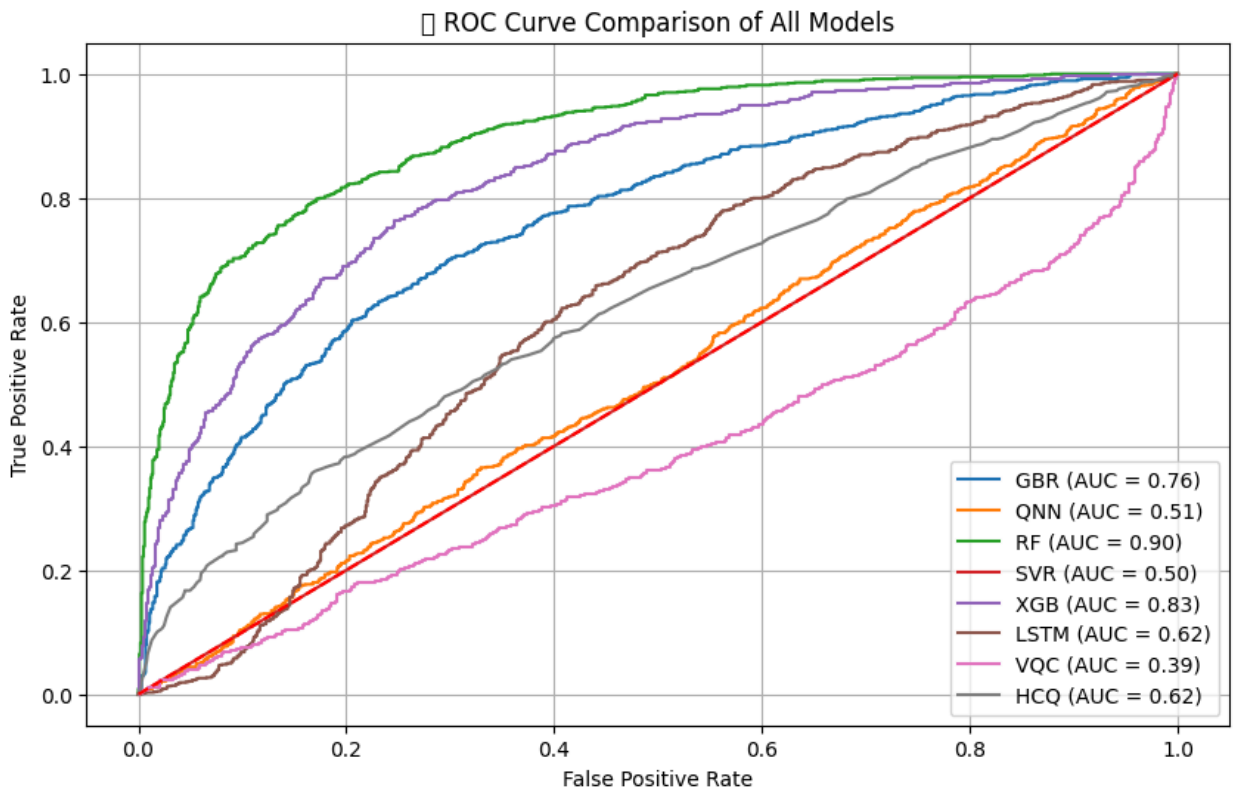
LSTM (Long Short-Term Memory): The residuals show a bimodal distribution with two peaks, suggesting the model might be struggling with certain subsets of the data. The residuals have a high spread, suggesting significant variability in errors.

VQC (Variational Quantum Circuit): The residuals show a unimodal distribution with a single peak. The distribution is slightly skewed to the right, indicating a tendency to underestimate. The residuals have a similar spread to QNN, RF, and XGB.

HCQ (Hybrid Classical-Quantum): The residuals show a bimodal distribution with two peaks, indicating potential issues with the model's fit. The residuals have a high spread, suggesting significant variability in errors.

GBR, SVR, LSTM, and HCQ models show bimodal residual distributions, indicating potential issues with model fit or data subsets. QNN, RF, XGB, and VQC models show unimodal residual distributions, but with varying degrees of skew and spread. Most models show a slight right skew, indicating a tendency to underestimate. SVR, LSTM, and HCQ models have a higher spread in their residuals, suggesting higher variability in errors.

In conclusion, the residual distributions reveal potential issues with model fit, underestimation bias, and error variability for different models. The bimodal distributions for GBR, SVR, LSTM, and HCQ suggest potential model fit issues, while the right skew in most models indicates an underestimation bias. The high spread in residuals for SVR, LSTM, and HCQ suggests higher error variability and potential unreliability.

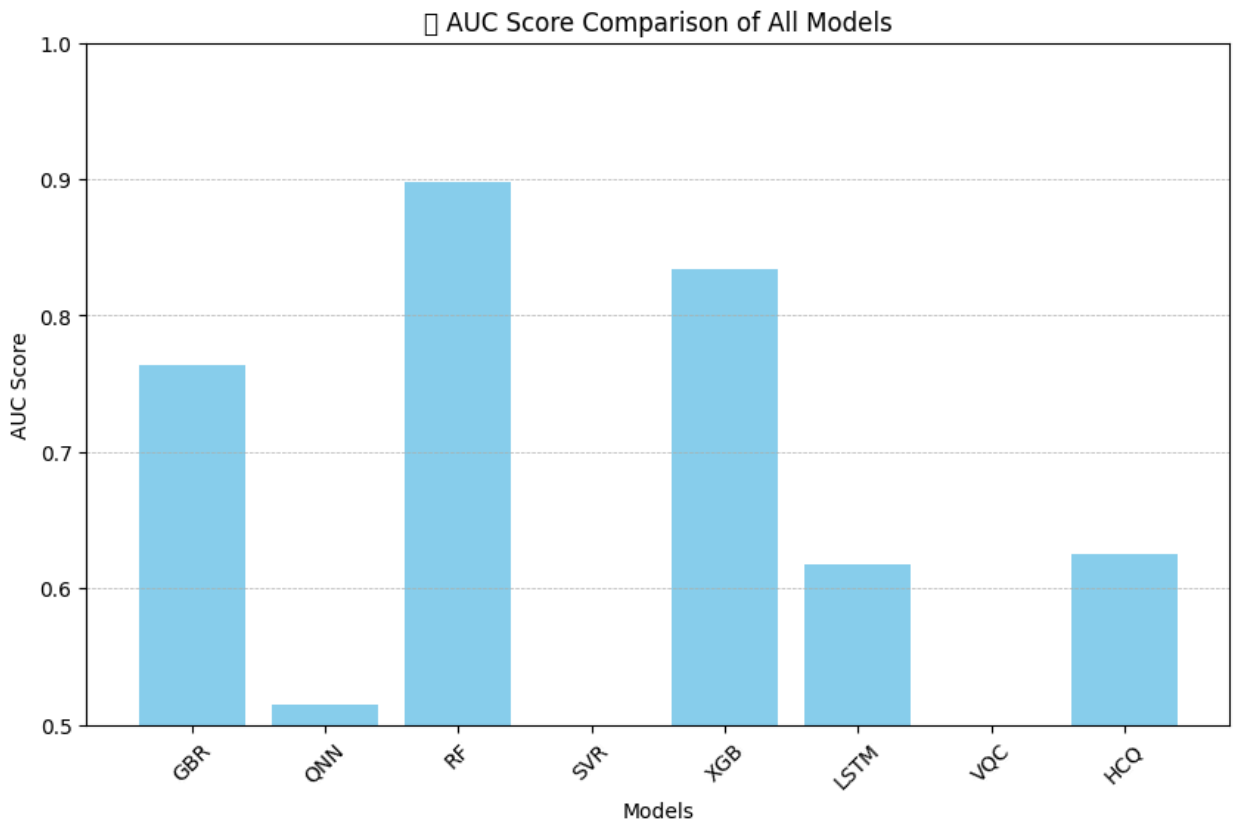


The ROC curve shows the trade-off between the true positive rate (TPR) and the false positive rate (FPR) for different threshold values. A good model aims to maximize TPR while minimizing FPR, resulting in a curve that is closer to the top-left corner. The AUC is a measure of the model's ability to distinguish between positive and negative classes. A higher AUC indicates better performance.

Model Performance Ranking (Based on AUC):

- Best: RF (AUC = 0.90)
- Good: XGB (AUC = 0.83)
- Decent: GBR (AUC = 0.76)
- Poor: LSTM (AUC = 0.62), HCQ (AUC = 0.62)
- Worst: QNN (AUC = 0.51), SVR (AUC = 0.50), VQC (AUC = 0.39)

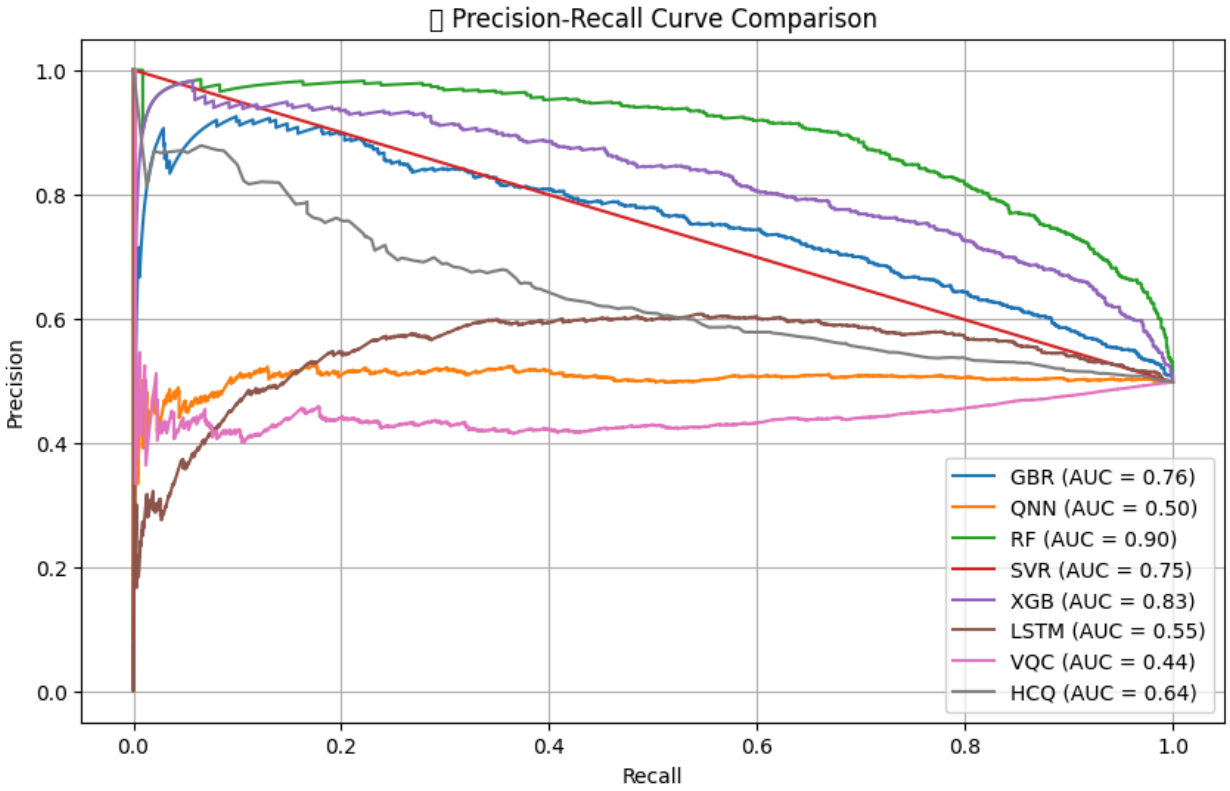
In conclusion, the ROC curve comparison shows that the Random Forest (RF) model is the best performer, followed by XGB and GBR. The LSTM and HCQ models have poor performance, while the QNN, SVR, and VQC models have very poor performance, with the VQC model performing worse than random guessing. The mislabeling of the SVR line suggests potential data or labeling errors.



The graph compares the performance of multiple models based on their AUC scores. The goal is to determine which model performs best in terms of its ability to distinguish between positive and negative classes.

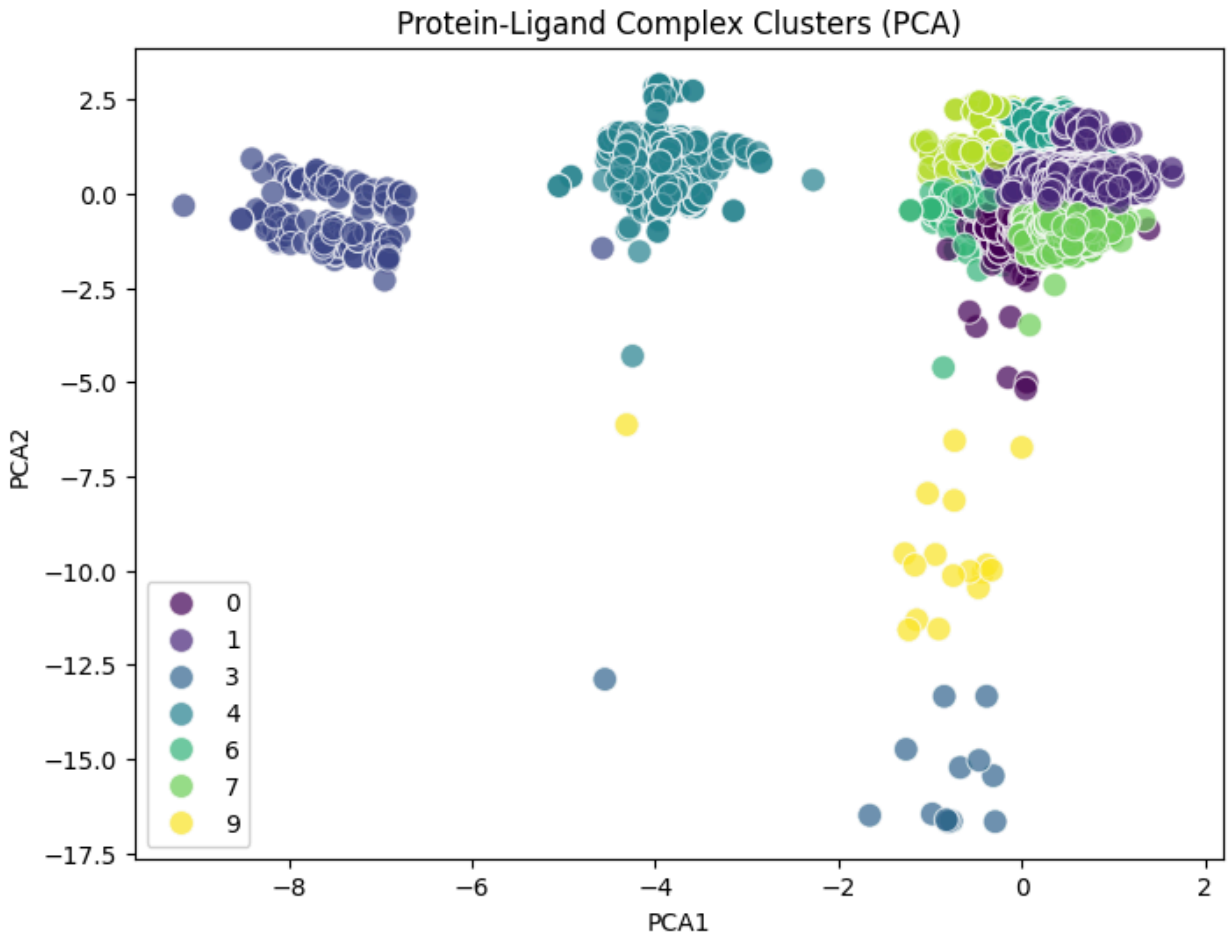
The bar graph clearly shows that the Random Forest (RF) model is the best performer, followed by XGB and GBR. The LSTM and HCQ models have poor performance, while the QNN, SVR, and VQC models have very poor performance, with the VQC model performing worse than random guessing. This suggests that the RF model is the most suitable choice for this

classification task.



The Precision-Recall (PR) curve shows the trade-off between precision and recall for different threshold values. A high area under the curve (AUC) represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. A high AUC indicates that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall).

In conclusion, the Precision-Recall curve comparison shows that the Random Forest (RF) model is the best performer, followed by XGB, GBR, and SVR. The HCQ and LSTM models have poor performance, while the QNN and VQC models have very poor performance, with the VQC model having the poorest precision-recall trade-off. This suggests that the RF model is the most suitable choice for this classification task.

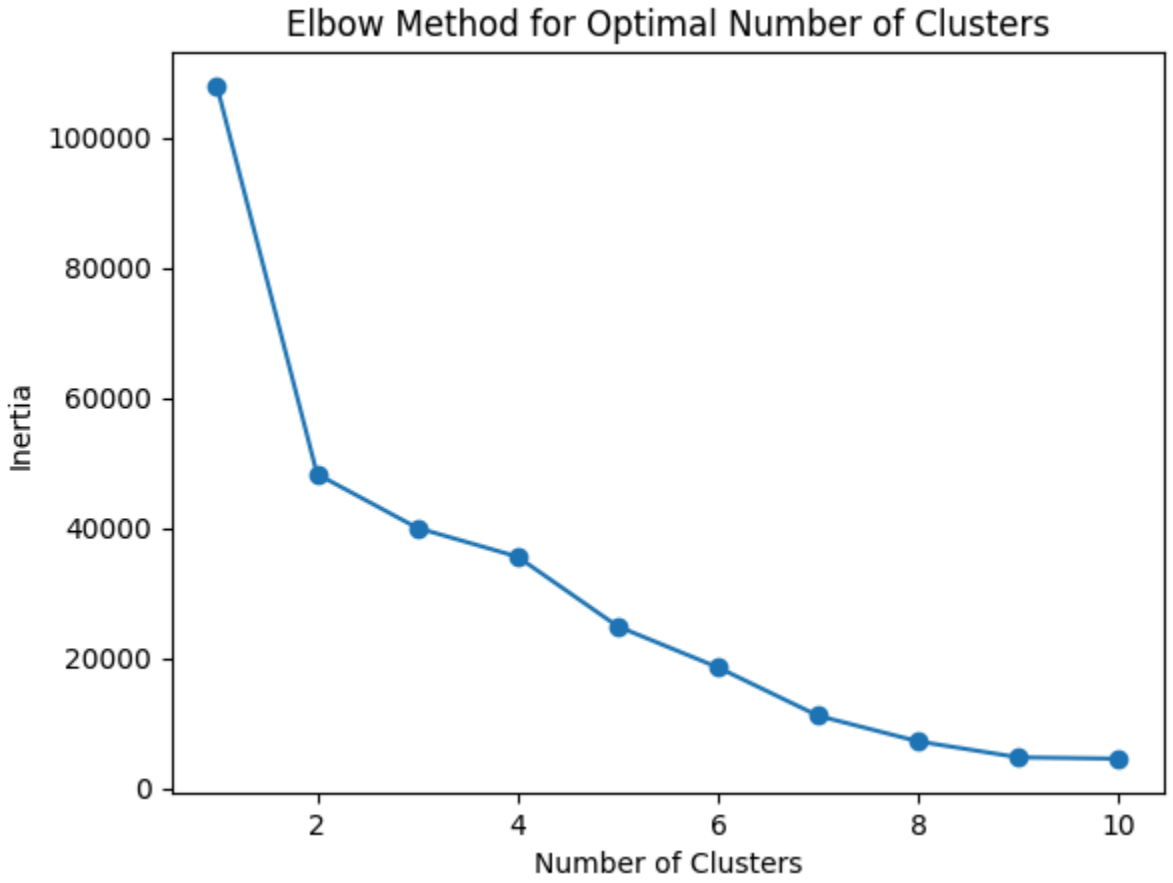


PCA has successfully reduced the high-dimensional data of protein-ligand complexes into a 2D representation, making it easier to visualize and analyze. The plot shows that the clustering algorithm has identified distinct clusters in the data. The different colors represent these clusters. Cluster 0 (Dark Purple) forms a tight, well-defined cluster in the top-left region. Cluster 1 (Grey) also forms a relatively tight cluster, but is slightly more spread out than cluster 0. Cluster 3 (Dark Blue) appears to be a small cluster near the center. Cluster 4 (Green) forms a distinct cluster in the upper-center region. Cluster 6 (Light Green) forms a cluster overlapping with cluster 4, but slightly more spread out. Cluster 7 (Yellow-Green) forms a cluster in the upper-right region. Cluster 9 (Yellow) forms a distinct cluster extending downwards along the PCA2 axis.

The fact that clusters are well-separated in this 2D space suggests that PCA1 and PCA2 have effectively captured the most significant variance in the original data, allowing for meaningful clustering. Clusters 4 and 6 are relatively close, suggesting they might share some similarities. Clusters 7 and 9 are distinct but show a gradient along the PCA2 axis, indicating a potential gradual change in properties.

The distinct clusters likely correspond to meaningful differences in the protein-ligand complexes. Complexes in different clusters might have different binding strengths. The way the ligand binds to the protein might vary across clusters. Complexes in different clusters might have distinct structural characteristics.

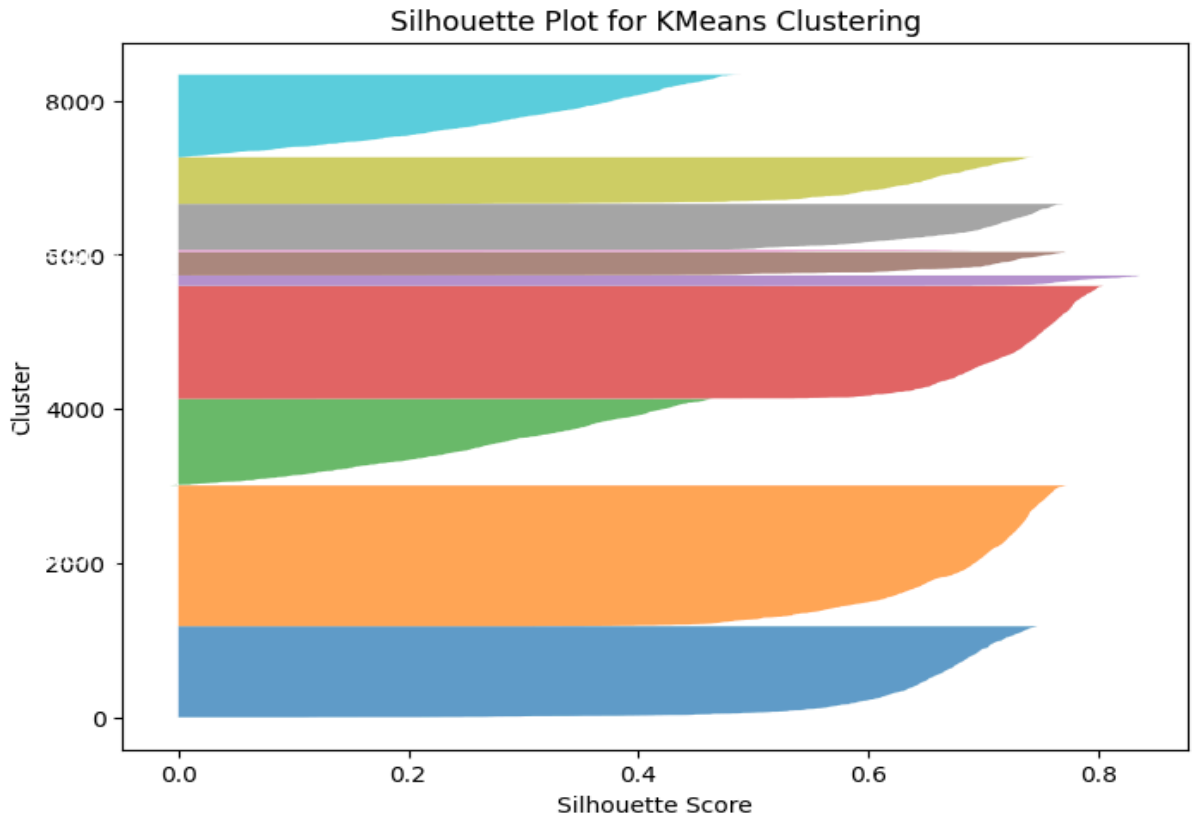
The graph effectively visualizes the clustering of protein-ligand complexes in a reduced-dimensional space using PCA. The distinct clusters suggest that the data can be meaningfully categorized based on underlying features captured by the principal components. This visualization can be a valuable tool for understanding the relationships between different complexes and identifying potential patterns or trends related to their biological activity. Further analysis of the features contributing to PCA1 and PCA2 could provide deeper insights into the specific differences between the clusters.



The Elbow Method aims to find the "elbow" point in the graph, where the rate of decrease in inertia significantly slows down. This point is considered the optimal number of clusters because adding more clusters beyond this point provides diminishing returns in terms of reducing inertia. As the number of clusters increases, the inertia decreases. This is expected because with more clusters, the data points within each cluster are closer to their respective centroids, resulting in lower within-cluster variance.

There's a sharp drop in inertia from 1 to 2 clusters. The rate of decrease in inertia slows down significantly after 2 clusters. The "elbow" point appears to be at 2 clusters. Beyond 2 clusters, the decrease in inertia is less substantial. Based on the Elbow Method, the optimal number of clusters for this dataset is likely 2.

The Elbow Method suggests that 2 clusters is the optimal number for this dataset. This indicates that the data can be effectively partitioned into two distinct groups. Using more clusters would not provide a significant improvement in the clustering quality and might lead to overfitting or less meaningful clusters. The Elbow Method is a useful technique for determining the appropriate number of clusters for algorithms like K-Means, helping to find a balance between model complexity and data representation.



Each cluster is represented by a "blade" or shape that extends along the silhouette score axis. The width of the blade represents the number of data points in the cluster. The position of the blade along the x-axis represents the silhouette scores of the data points within that cluster. Ideally, we want blades that are wide (indicating many data points) and extend far to the right (indicating high silhouette scores).

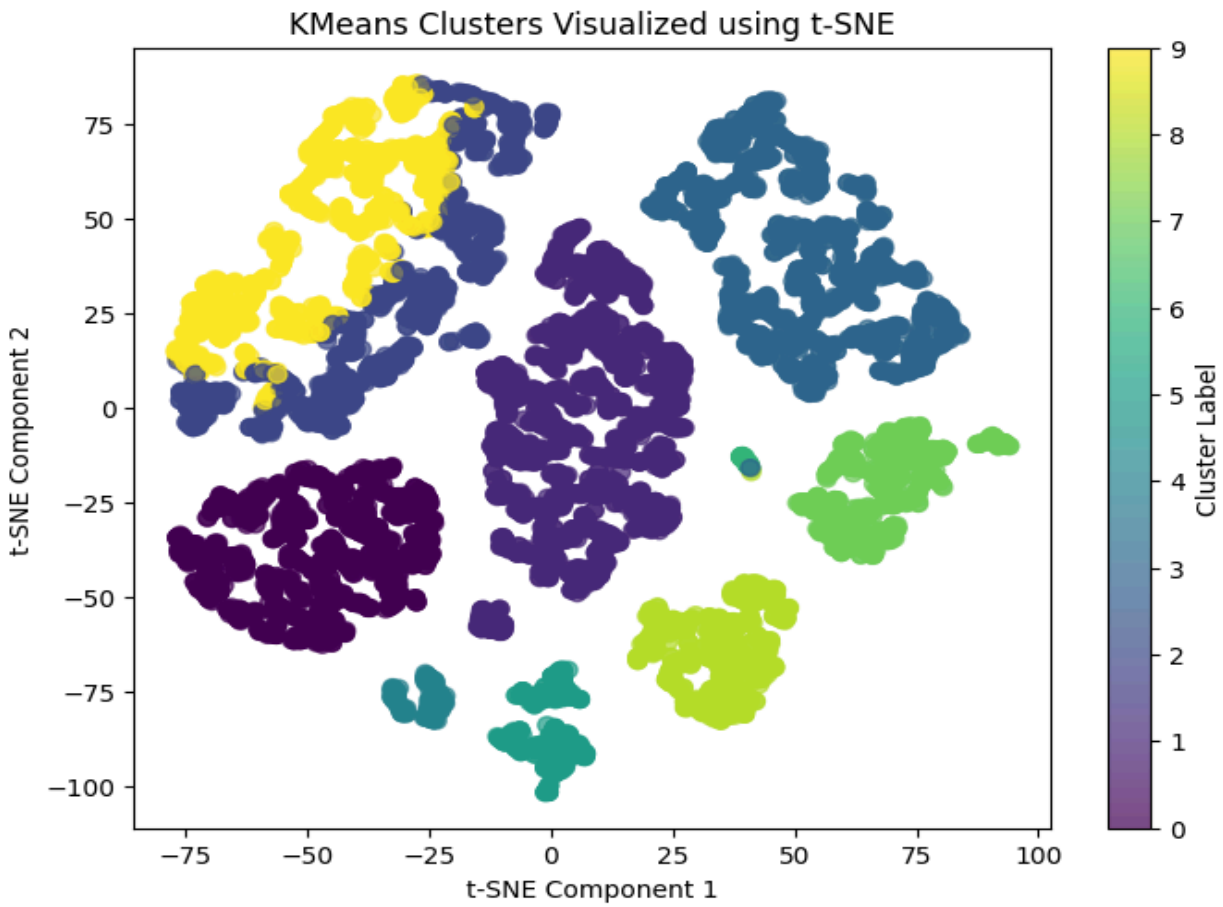
The average silhouette score for all data points is not explicitly shown but can be inferred from the distribution of blades. A higher average silhouette score indicates better clustering. The silhouette scores for most clusters range from approximately 0.0 to 0.8. This suggests that while some data points are well-clustered, others are closer to the decision boundaries or might be misclassified.

The blades have varying lengths, indicating that the data points within some clusters are more consistently well-clustered than others. Cluster 0 has a relatively high silhouette score (extending towards 0.8), suggesting good clustering. Cluster 1 also shows good clustering, with a reasonable silhouette score range. Clusters 2 and 3 have relatively lower silhouette scores, suggesting some data points might be misclassified or close to the decision boundary.

The blades have varying widths, indicating that the clusters have different sizes. Clusters 0, 1, and 4 appear to be relatively larger, containing more data points. Clusters 2, 3, 5, 6, 7, and 8 are smaller. Clusters 2 and 3 have some data points with low silhouette scores, suggesting potential issues with the cluster assignments.The varying blade lengths and silhouette scores suggest potential overlap or ambiguity between some clusters.

The Silhouette Plot indicates that the KMeans clustering has produced clusters with varying quality. Some clusters (like 0 and 1) show good consistency and separation, while others (like 2 and 3) have lower silhouette scores, suggesting potential misclassifications. The varying cluster sizes indicate that the data is not evenly distributed across the clusters.

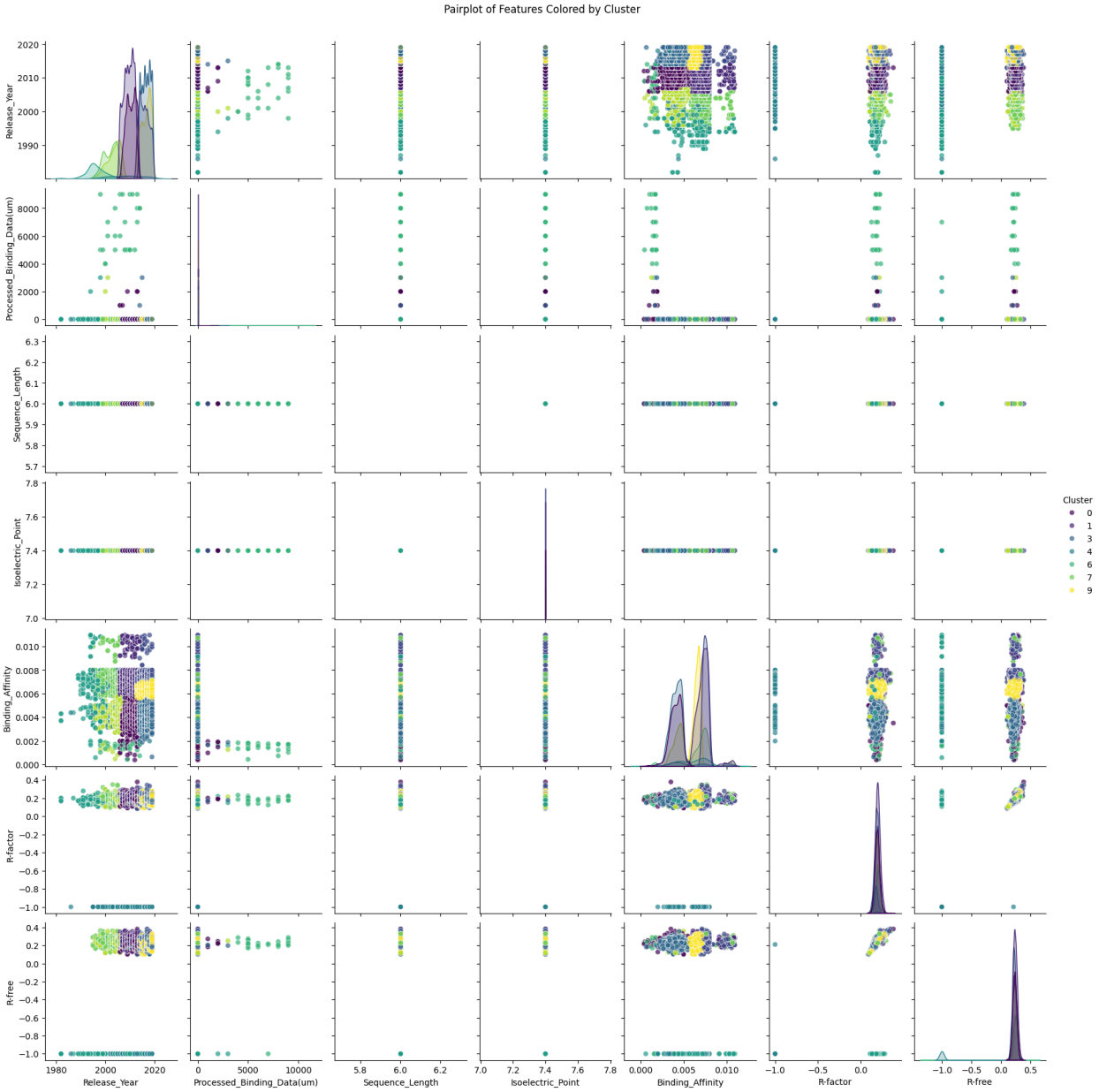
This plot can be used to assess the effectiveness of the KMeans algorithm and to potentially refine the clustering process. For example, one might consider adjusting the number of clusters or using a different clustering algorithm to improve the overall quality of the clusters. An average silhouette score (if available) would provide a more direct measure of the overall clustering quality.



t-SNE has successfully reduced the high-dimensional data into a 2D representation, making it easier to visualize and analyze the KMeans clusters. The plot shows that the KMeans algorithm has identified distinct clusters, which are well-separated in the t-SNE space.

The different colors represent distinct clusters, suggesting that KMeans has effectively grouped similar data points together. The clusters are relatively well-separated, with clear boundaries between them. This indicates that the KMeans algorithm has found meaningful patterns in the data. t-SNE is particularly good at preserving local relationships in the data, which is evident in the non-linear shapes of the clusters. This suggests that the underlying data might have complex, non-linear structures. The clear separation of clusters in the t-SNE plot suggests that KMeans has performed well in grouping the data points.

The graph effectively visualizes the results of KMeans clustering using t-SNE dimensionality reduction. The distinct and well-separated clusters indicate that KMeans has successfully grouped the data points based on their similarities. This visualization can be a valuable tool for understanding the relationships between different data points and identifying potential patterns or trends within the data. The clear separation of clusters suggests that the data exhibits strong clustering patterns, and further analysis of the features contributing to these clusters could reveal important insights about the underlying structure of the data.

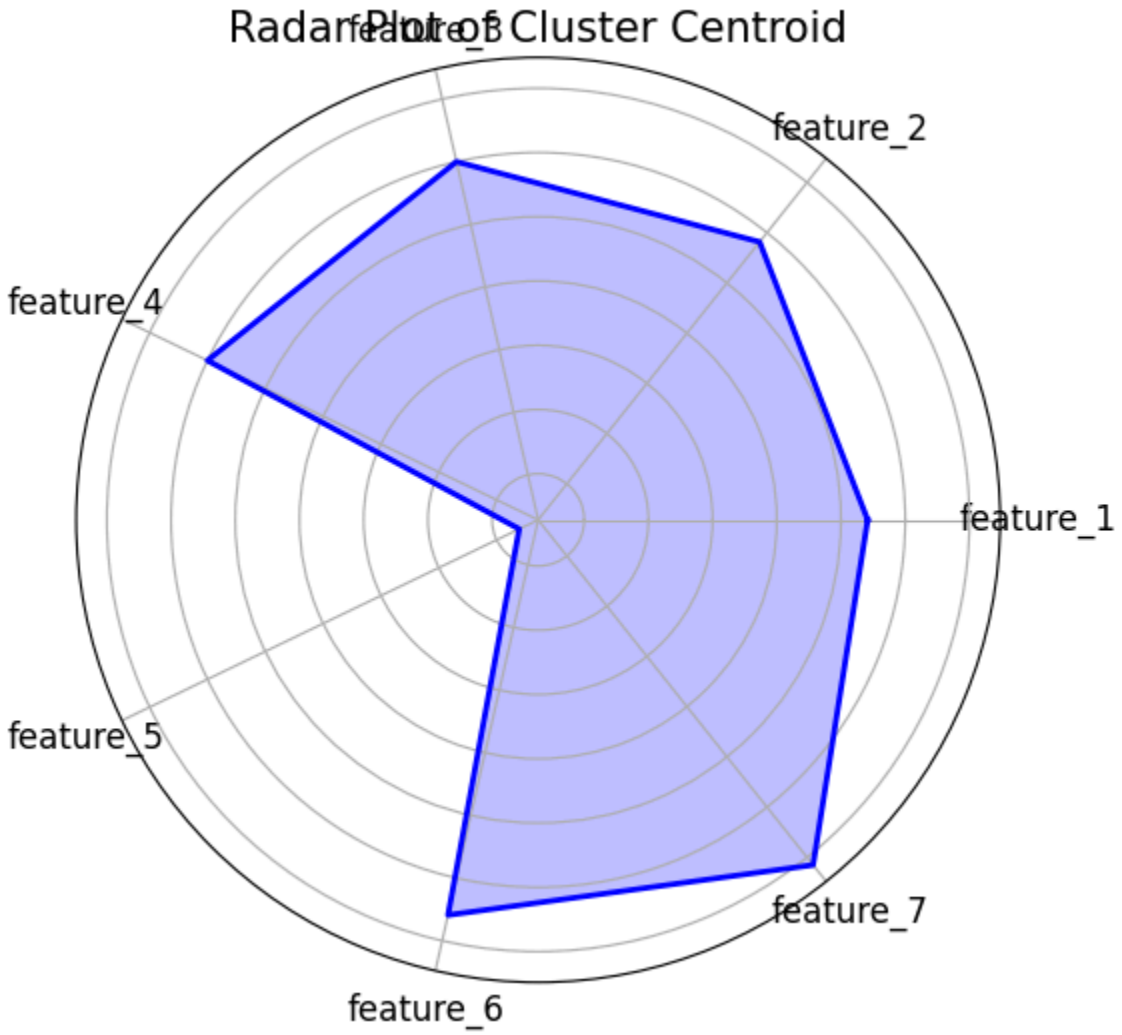


This image shows a pairplot, a matrix of scatter plots and histograms, visualizing the relationships between multiple features in a dataset, colored by cluster labels. Let's break it down:

There seems to be some separation of clusters in the "Release Year" vs. "Binding Affinity" plot. Different clusters might correspond to different trends in binding affinity over time. There's likely a strong relationship between these two, as they both relate to binding strength. The plot might show a correlation. The relationship here is not immediately clear and needs closer inspection.

These two features are likely correlated, as they both relate to data quality. The plot should show a clear relationship. Some clusters might overlap in certain feature combinations, indicating that they are not perfectly separable based on the visualized features.

This pairplot provides a valuable overview of the relationships between features in the dataset, colored by cluster labels. It allows us to visually assess cluster separation, identify feature relationships, and understand feature distributions. Further in-depth analysis of the plots is necessary to draw more concrete conclusions about the data and the clustering results.

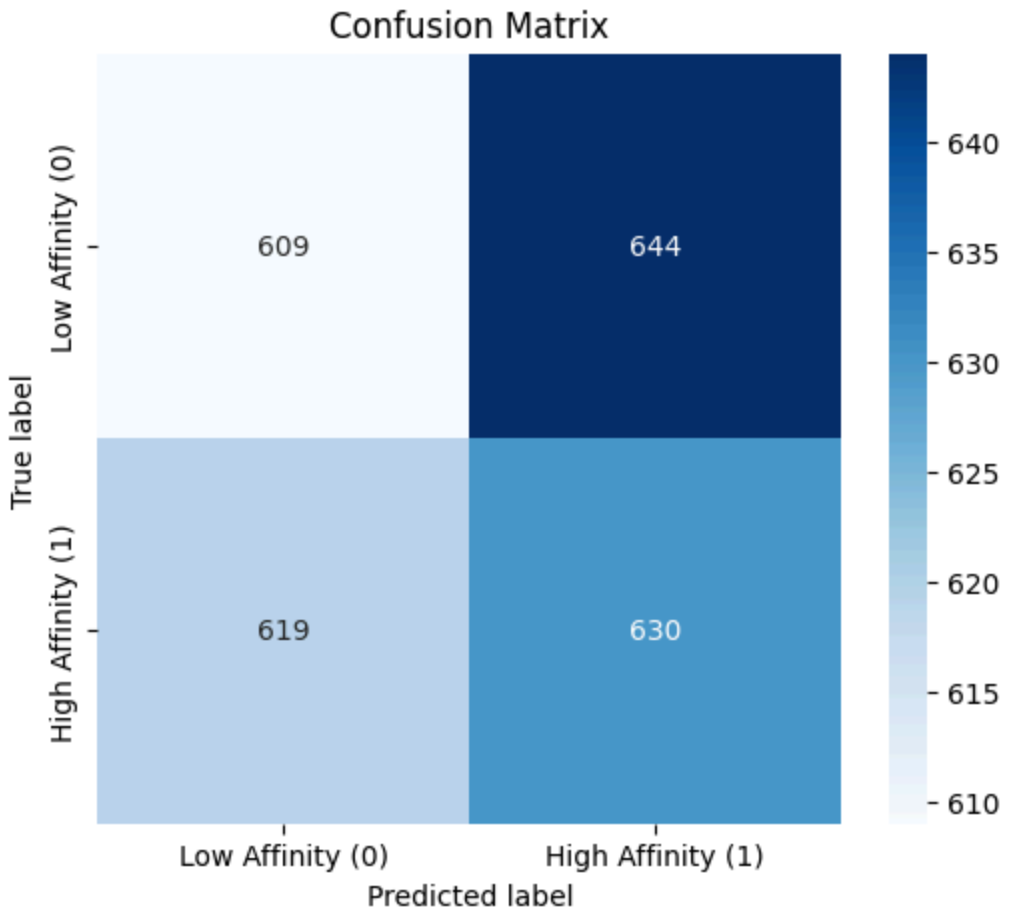


This is a radar plot (also known as a spider chart or star chart) representing the centroid of a cluster in a multi-dimensional feature space. The blue line connects the values of the cluster centroid for each feature. The blue shaded area represents the "area" covered by the cluster centroid's feature values.

The plot shows that "feature_4" has the highest value for this cluster's centroid, indicating it's a prominent characteristic of the cluster. "feature_1" and "feature_2" have relatively high values, suggesting they are also important for defining this cluster. "feature_6" and "feature_7" have moderately high values. "feature_5" has the lowest value, suggesting it's less influential in defining this cluster.

The high value of "feature_4" suggests that this feature might be a distinguishing characteristic of the cluster. The plot can help in understanding the meaning or interpretation of the cluster based on the features that have high values. The plot provides a visual representation of feature importance for a given cluster. Features with higher values in the centroid are likely more important for defining the cluster. If multiple radar plots are available (one for each cluster), they can be compared to understand the differences between clusters in terms of their feature values.

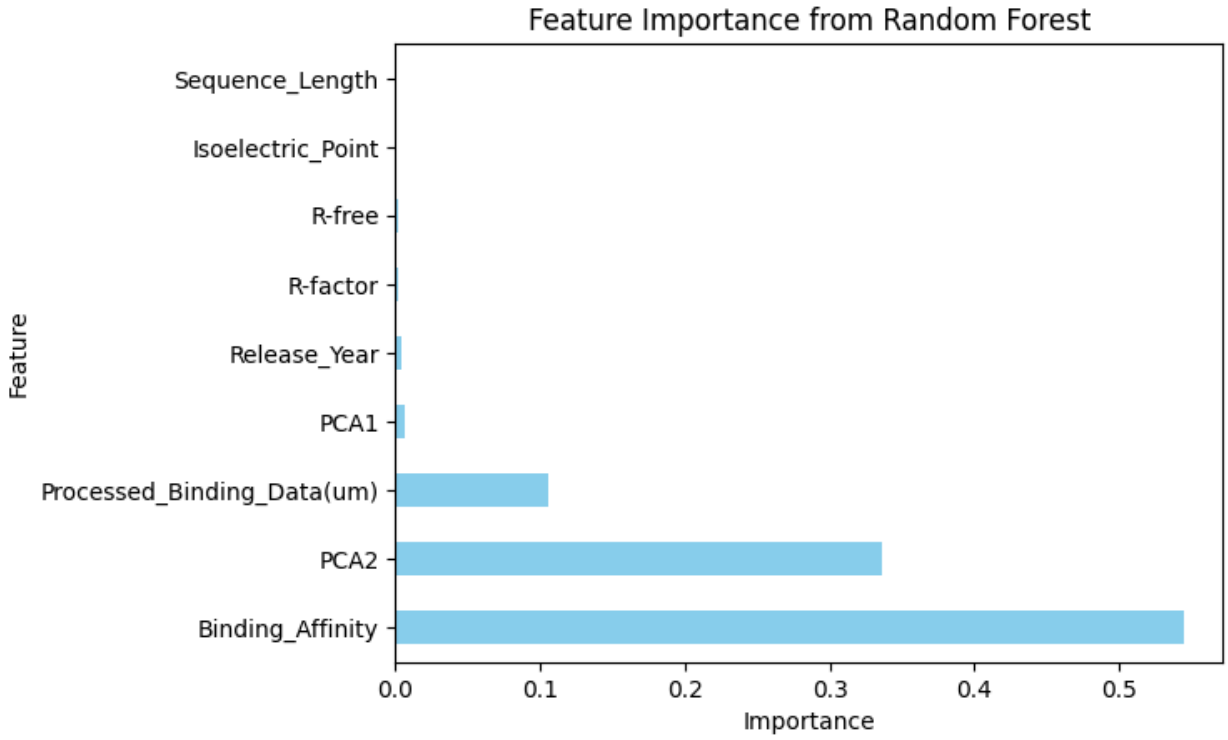
This radar plot provides a concise and intuitive way to visualize the centroid of a cluster across multiple features. It highlights the relative importance of different features in characterizing the cluster and helps in understanding the cluster's properties. The plot is particularly useful for comparing clusters based on their feature centroids and for identifying the most important features that distinguish each cluster.



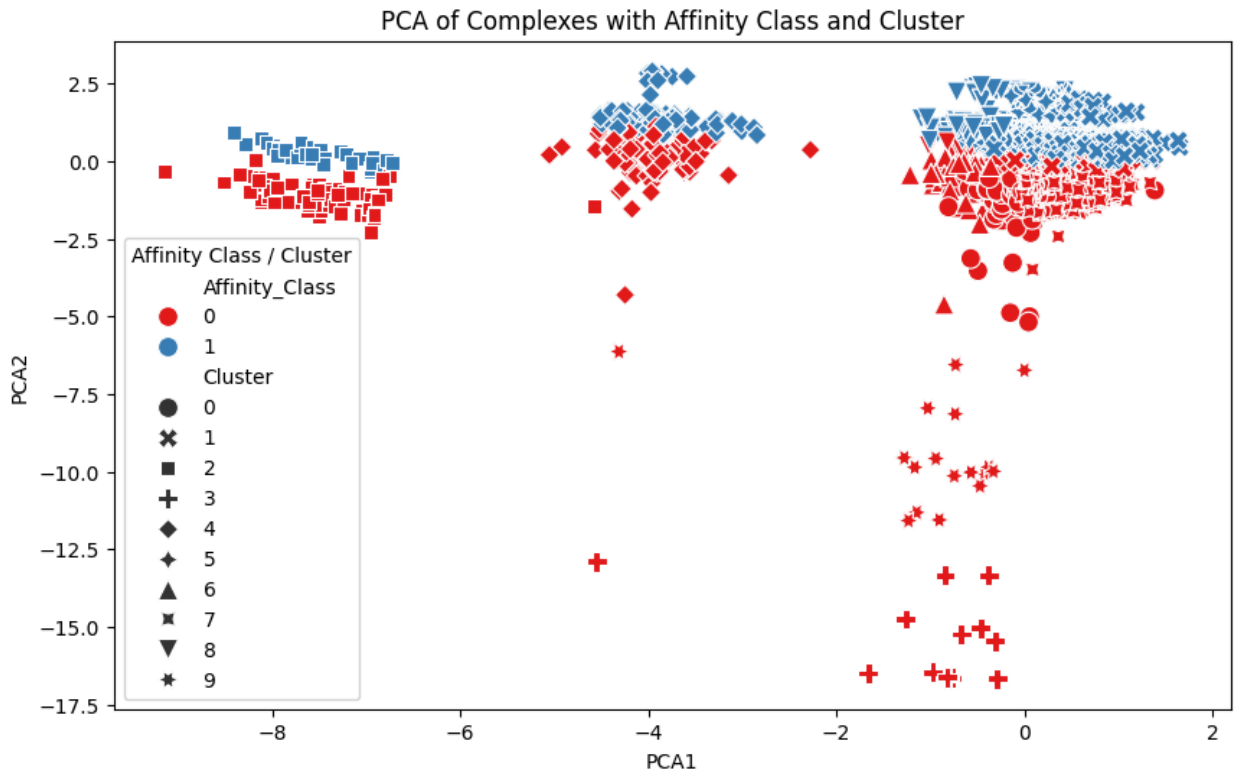
True Negatives (TN). The model correctly predicted 609 instances as "Low Affinity (0)" when they were actually "Low Affinity (0)". False Positives (FP). The model incorrectly predicted 644 instances as "High Affinity (1)" when they were actually "Low Affinity (0)". False Negatives (FN). The model incorrectly predicted 619 instances as "Low Affinity (0)" when they were actually "High Affinity (1)". True Positives (TP). The model correctly predicted 630 instances as "High Affinity (1)" when they were actually "High Affinity (1)".

(TN + TP) / (TN + FP + FN + TP) = (609 + 630) / (609 + 644 + 619 + 630) = 1239 / 2502 ≈ 0.495 or 49.5%. This means the model is only accurate about 49.5% of the time, which is close to random guessing.

The model has a significant number of false positives (644), indicating it often predicts "High Affinity (1)" when it should be "Low Affinity (0)". The model also has a significant number of false negatives (619), indicating it often predicts "Low Affinity (0)" when it should be "High Affinity (1)". The matrix shows a relatively balanced dataset, with a similar number of instances in each class (approximately 1250 each). Therefore, class imbalance is not a major concern here.

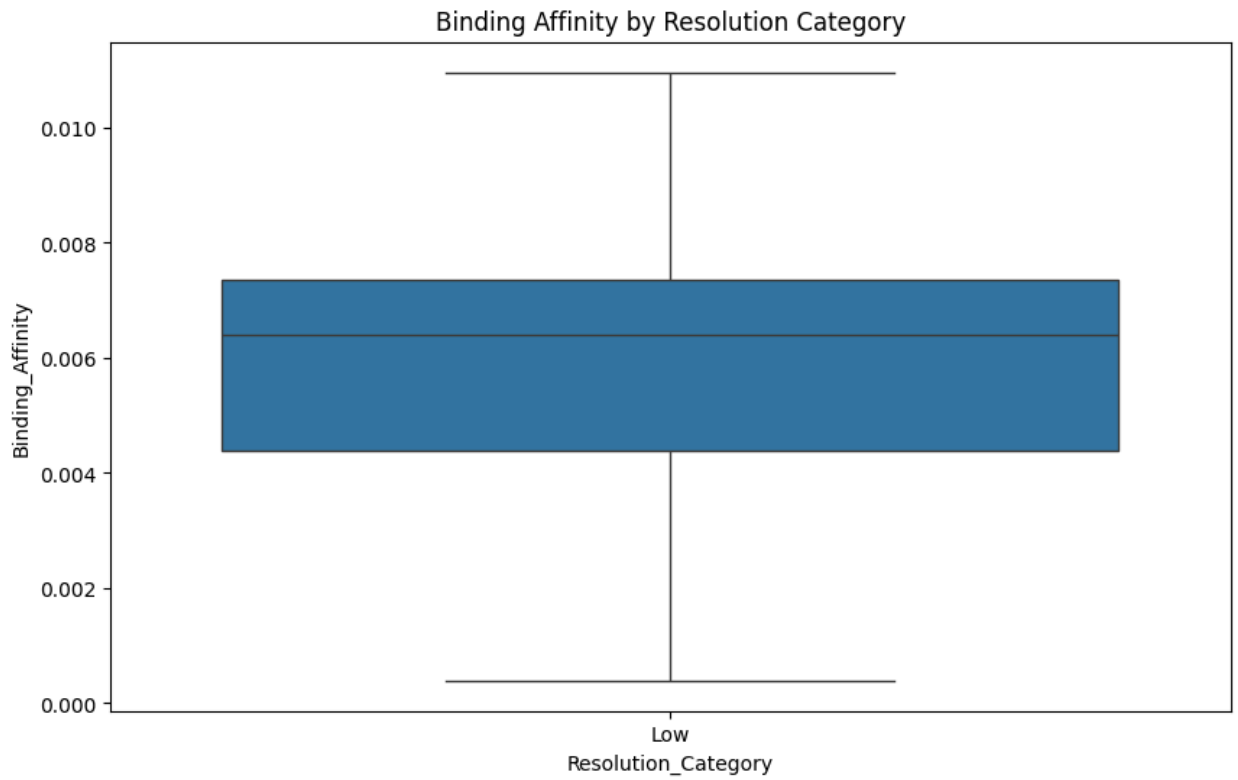


The chart shows that "Binding_Affinity" is the most important feature in the Random Forest model, followed by "PCA2" and "Processed_Binding_Data(um)". The other features have very low importance and might be considered for removal. This information can be valuable for understanding the relationships between features and the target variable, as well as for improving the model's performance by focusing on the most relevant features.



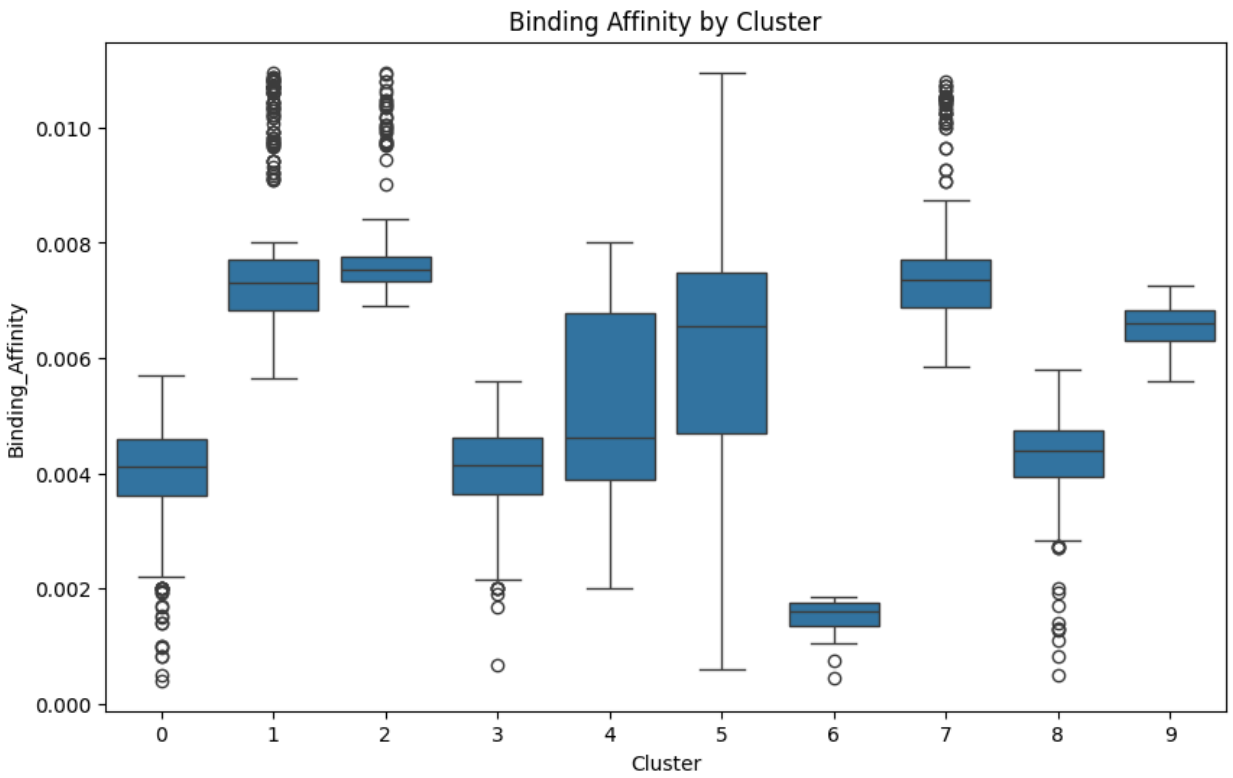
The partial separation of affinity classes suggests that there are some underlying differences in the complexes that contribute to their different affinities. The distinct clusters likely correspond to different structural or functional properties of the complexes.

The graph shows that PCA has successfully reduced the dimensionality of the complex data and revealed some patterns related to affinity classes and clusters. While there is some separation between the affinity classes, there is also significant overlap, suggesting that other factors might be involved in determining affinity. The clusters, represented by different shapes, show distinct distributions and are not perfectly aligned with the affinity classes. This visualization can be a valuable tool for understanding the relationships between different complexes and identifying potential patterns or trends related to their affinity and structural properties. Further analysis of the features contributing to PCA1 and PCA2 could provide deeper insights into the specific differences between the complexes.



The boxplot shows the distribution of binding affinity values specifically for the "Low" resolution category. The median binding affinity for low resolution is around 0.006. The IQR (the height of the box) shows the spread of the middle 50% of the data. The IQR for low resolution is relatively narrow, indicating that the middle 50% of the binding affinity values are clustered closely together. The whiskers show the full range of binding affinity values for low resolution, excluding outliers. The range is from approximately 0.0002 to 0.011. The boxplot is relatively symmetrical, suggesting that the distribution of binding affinity values for low resolution is approximately symmetrical. There are no outliers plotted outside the whiskers, indicating that there are no extreme values in the binding affinity data for low resolution.

The boxplot shows the distribution of binding affinity values for the "Low" resolution category. The median binding affinity is around 0.006, and the data is relatively clustered, with no outliers. This visualization provides a concise summary of the binding affinity data for low resolution and can be used to compare it with other resolution categories if they were included in the plot.



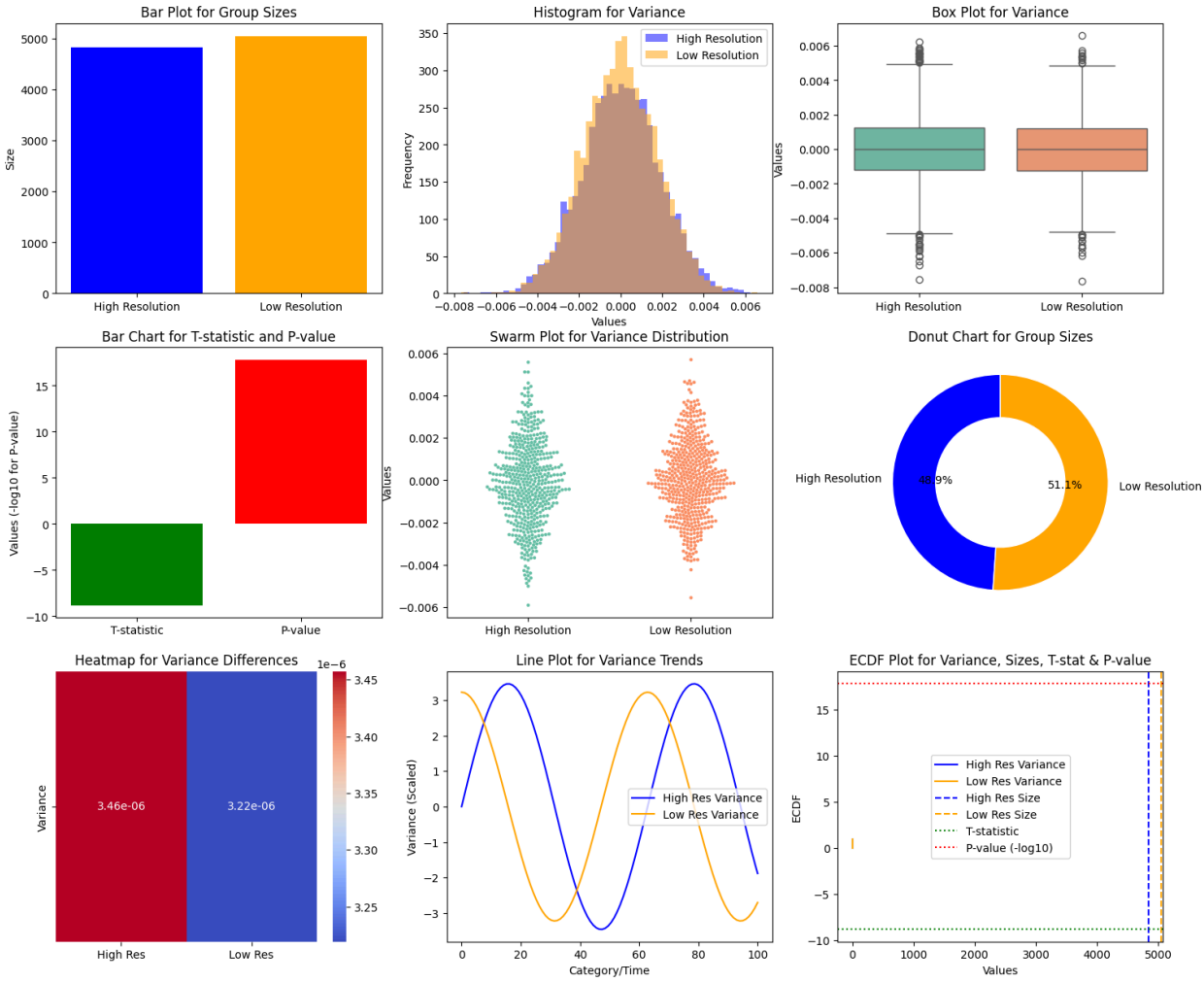
The boxplot shows the distribution of binding affinity values for each cluster (0 to 9). Cluster 6 has the lowest median binding affinity, close to 0.001. Clusters 2, 5, 7 these clusters have relatively high median binding affinities, around 0.007 to 0.008. Clusters 0, 1, 3, 4, 8, 9 these clusters have median binding affinities in the range of 0.004 to 0.007. Cluster 4 has the largest IQR, indicating a wide spread of binding affinity values within the middle 50% of the data. Cluster 6 has the smallest IQR, indicating a narrow spread of binding affinity values. Cluster 5 has the widest range of binding affinity values (excluding outliers), from approximately 0.0005 to 0.011. Cluster 6 has the narrowest range of binding affinity values (excluding outliers), from approximately 0.0005 to 0.002.

Clusters 1, 2, 4, 5, 7, 8 these clusters have a significant number of outliers, indicating extreme binding affinity values. Clusters 0, 3, 6, 9 these clusters have relatively fewer outliers. Clusters 1, 2, 4, 5, 7, 8 these clusters show some degree of asymmetry, with longer whiskers on one side and a larger number of outliers. Clusters 0, 3, 6, 9 these clusters are relatively symmetrical.

The differences in median, IQR, range, and outliers suggest that the clusters have distinct binding affinity profiles. The differences in binding affinity across clusters might reflect different structural or functional properties of the complexes within each cluster.

The boxplot shows that the clusters have distinct binding affinity distributions, with differences in median, IQR, range, and outliers. This suggests that the clusters represent groups of complexes with different binding affinity characteristics. The differences might reflect underlying structural

or functional variations within the complexes. Further analysis of the features contributing to the cluster differences could provide deeper insights into the specific properties of each cluster.



Bar Plot for Group Sizes shows the sizes of "High Resolution" and "Low Resolution" groups. The "High Resolution" group has a significantly larger size (around 5000) compared to the "Low Resolution" group (around 3000). This indicates an imbalance in the data distribution between the two resolution categories.

Histogram for Variance the distributions of variance for both groups are approximately normal and centered around zero. The "Low Resolution" variance distribution has a slightly wider spread, suggesting higher variability.

Box Plot for Variance is similar to the histogram, the box plots show that the variance distributions are centered around zero for both groups. The "Low Resolution" group has a slightly wider IQR, confirming higher variability. There are a few outliers in both groups.

Bar Chart for T-statistic and P-value shows the T-statistic is negative, indicating a difference in means between the two groups. The P-value (represented as -log10) is high, suggesting the difference is statistically significant.

Swarm Plot for Variance Distribution provides a more detailed view of the variance distribution compared to the box plot. It confirms the higher variability in the "Low Resolution" group.

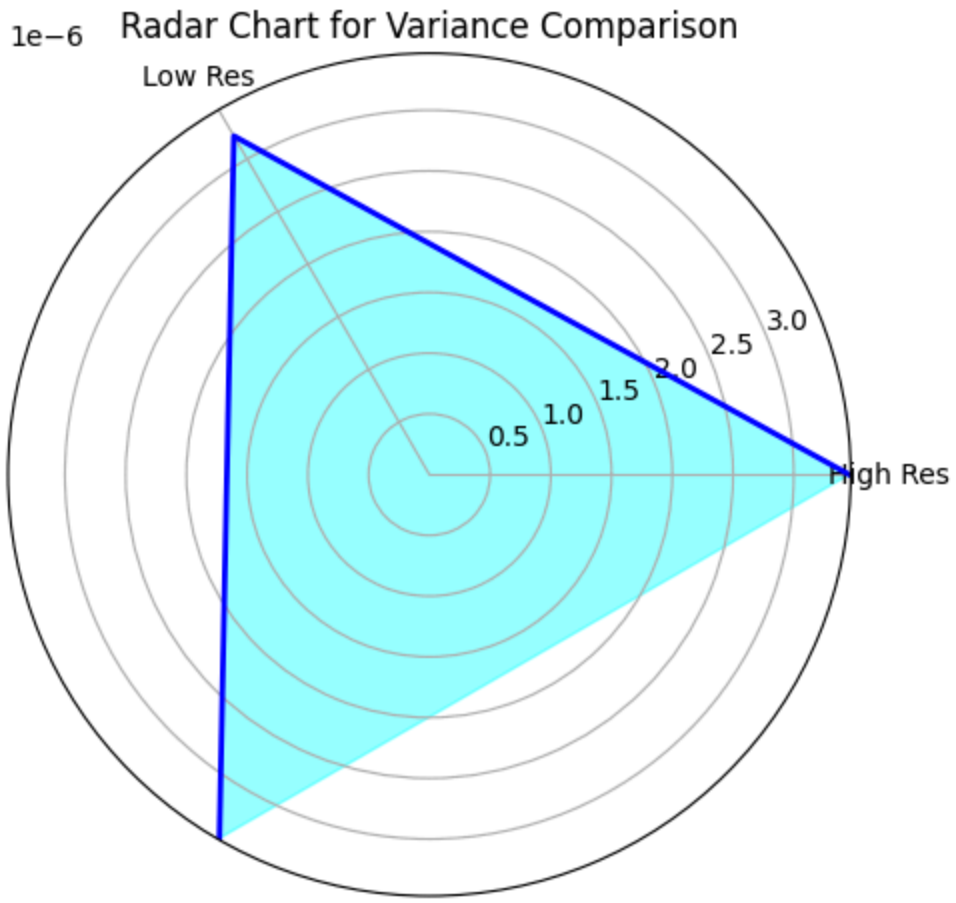
Donut Chart for Group Sizes reinforces the imbalance in group sizes, showing that "High Resolution" constitutes a significantly larger portion (51.1%) of the data compared to "Low Resolution" (9%).

Heatmap for Variance Differences the heatmap indicates a small but statistically significant difference in variance between the two groups. The "Low Resolution" group has a slightly higher variance.

Line Plot for Variance Trends the line plots suggest a cyclical pattern in variance over "Category/Time." The "Low Resolution" group shows a slightly higher magnitude of fluctuations.

ECDF Plot for Variance, Sizes, T-stat & P-value provides a comprehensive view of the distributions of all four variables. It confirms the differences in variance and size between the groups and highlights the statistical significance of the T-statistic and P-value.

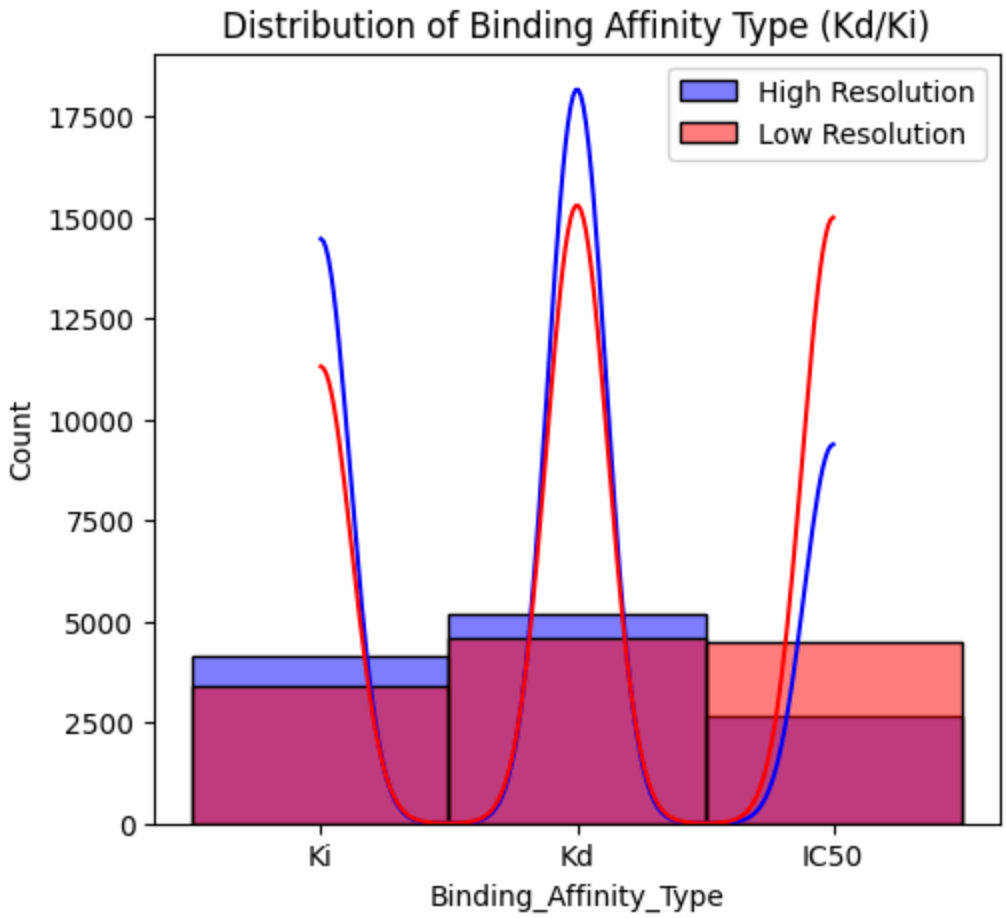
This visualization provides a comprehensive analysis of the variance differences between "High Resolution" and "Low Resolution" groups. It highlights the imbalance in group sizes, the slightly higher variability in the "Low Resolution" group, and the statistical significance of the variance difference. The cyclical pattern in variance trends over "Category/Time" is also noteworthy.



Radar chart (also known as a spider chart or star chart) is the chart indicating that the variance values are on the order of 1e-6 (1 x 10^-6). This is shown in the chart title and the values on the axes. The "High Res" category has a variance value around 2.5 x 10^-6 (based on the position of the blue line along the "High Res" axis). The "Low Res" category has a variance value around 3 x 10^-6 (based on the position of the blue line along the "Low Res" axis). The "Low Res" category has a slightly higher variance compared to the "High Res" category, as indicated by the larger value along the "Low Res" axis and the larger area covered by the shaded region towards "Low Res."

The higher variance in the "Low Res" category suggests that the data in this category is more spread out or has greater variability compared to the "High Res" category. The difference in variance might indicate differences in data quality or consistency between the two resolution categories. The higher variance in "Low Res" might warrant further investigation to understand the reasons for the increased variability and its potential impact on analysis or modeling.

The radar chart shows that the "Low Res" category has a slightly higher variance compared to the "High Res" category. This suggests that the data in the "Low Res" category is more variable or spread out. The magnitude of the variance is on the order of 1e-6.

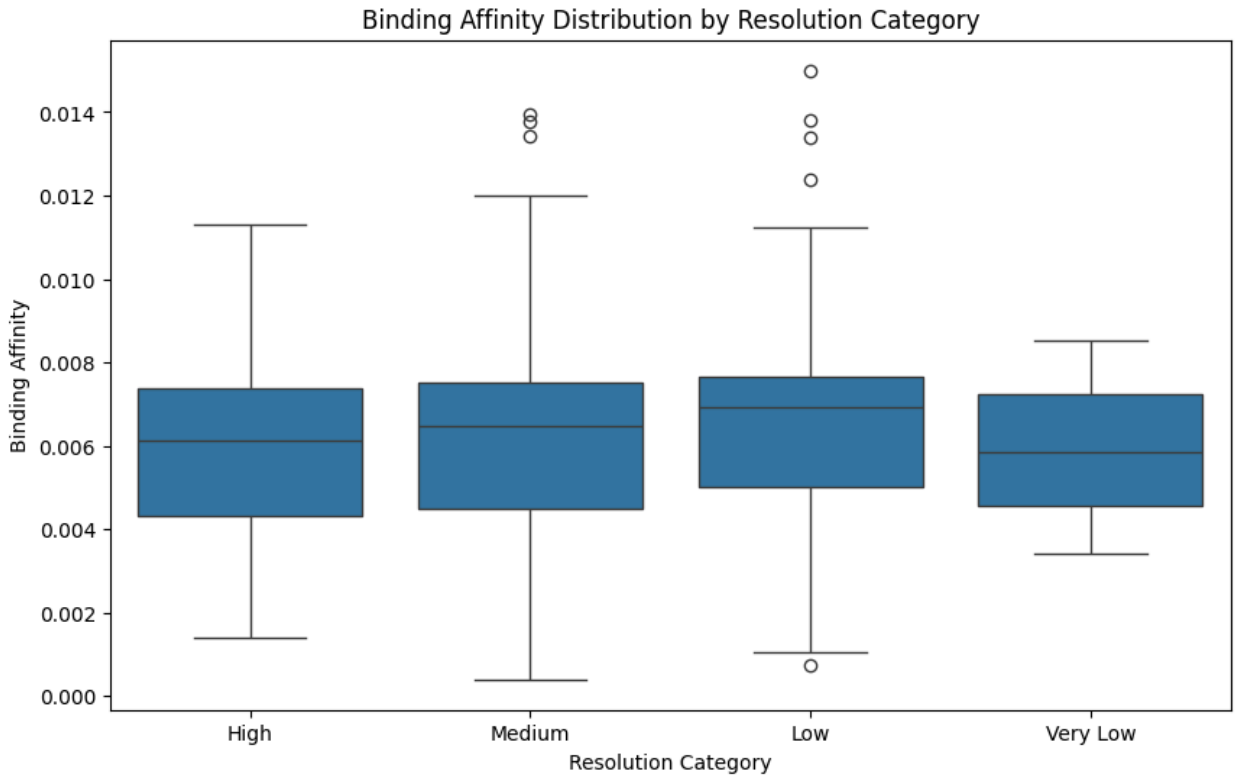


The graph compares the distribution of binding affinity types between "High Resolution" and "Low Resolution" data. Both "High Resolution" and "Low Resolution" data show a strong preference for Kd measurements. The counts for Kd are significantly higher than Ki or IC50. "High Resolution" has a slightly higher count of Ki compared to "Low Resolution". "Low Resolution" has a slightly higher count of IC50 compared to "High Resolution."

The distribution lines show a sharp peak for Kd in both categories, indicating a strong concentration of data points at Kd. The distribution lines show smaller peaks for Ki and IC50, indicating a lower concentration of data points. The strong preference for Kd measurements might indicate a bias in the data collection or reporting process.

The graph reveals the distribution of binding affinity types in the dataset and highlights the dominance of Kd measurements. The preference for Kd might reflect experimental design choices or the availability of data for different binding affinity types. The differences in Ki and IC50 counts between resolution categories might suggest variations in data quality or experimental procedures.

The graph shows that the dataset is heavily skewed towards Kd measurements, with significantly higher counts compared to Ki and IC50. The "High Resolution" data has a slightly higher count of Ki, while the "Low Resolution" data has a slightly higher count of IC50. This information can be useful for understanding the characteristics of the data and for considering potential biases or limitations in further analysis. The strong preference for Kd might warrant further investigation to understand the reasons for this bias and its potential impact on conclusions.

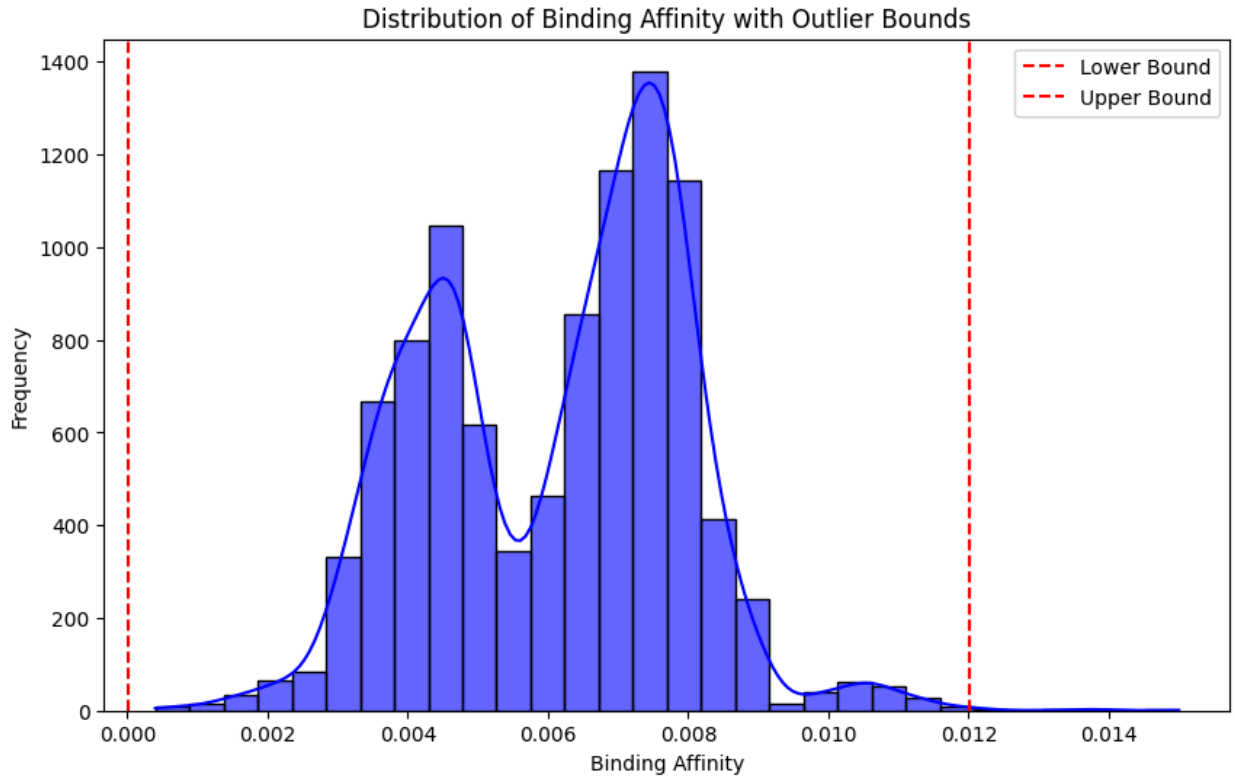


The boxplot shows the distribution of binding affinity values for each resolution category (High, Medium, Low, Very Low). Medium and Low these categories have the highest median binding affinities, around 0.0065 to 0.007. High and Very Low these categories have relatively lower median binding affinities, around 0.006.

The "Low" resolution category has the widest range of binding affinity values (excluding outliers), from approximately 0.0008 to 0.012. The "Very Low" resolution category has the narrowest range of binding affinity values (excluding outliers), from approximately 0.003 to 0.0085. Medium and Low these categories have the most outliers, indicating extreme binding affinity values. High and Very Low these categories have relatively fewer outliers.

The boxplot shows that the distribution of binding affinity varies across different resolution categories. The "Medium" and "Low" resolution categories have higher median binding affinities and more outliers compared to "High" and "Very Low" categories. The "Low" resolution category

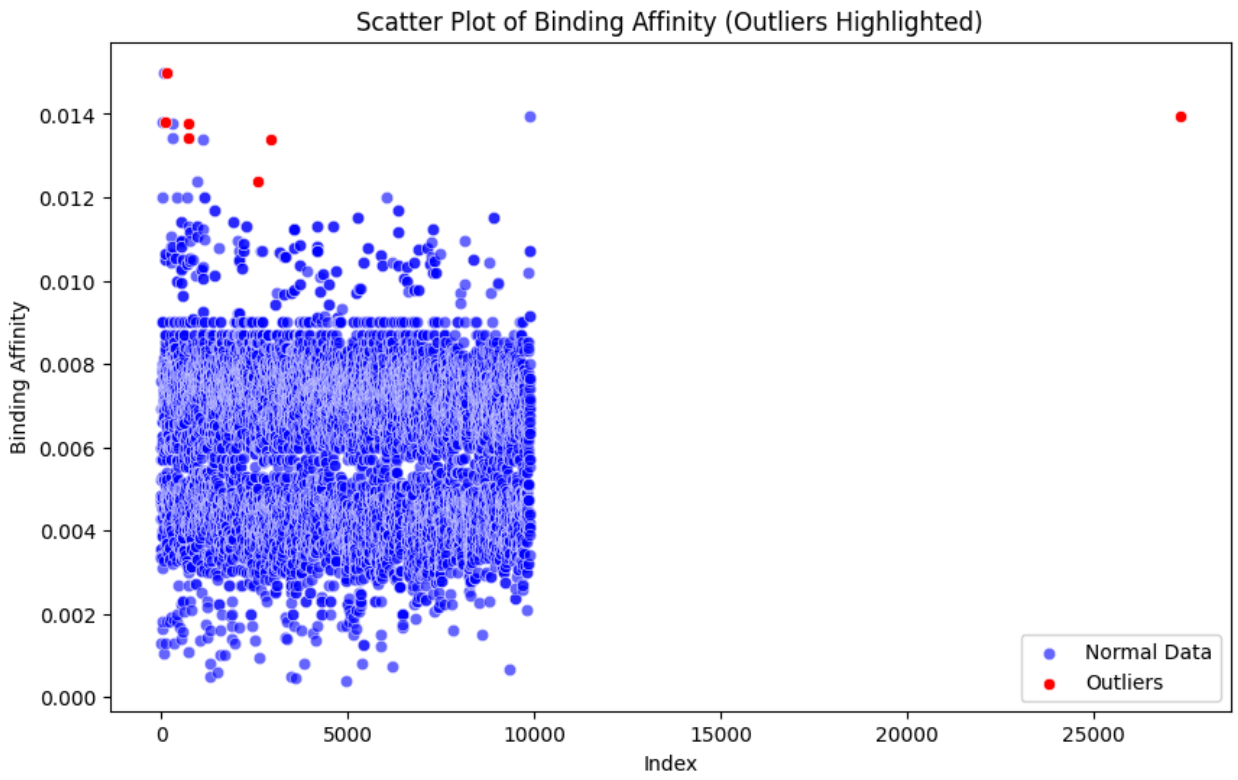
also exhibits the widest range and IQR, indicating higher variability in binding affinity values. These differences might reflect variations in data quality or experimental procedures associated with different resolution levels.



The distribution of binding affinity values appears to be bimodal, with two distinct peaks. This suggests that the data might be composed of two subgroups or populations with different binding affinity characteristics. The first peak is centered around 0.0045. The second peak is centered around 0.008. The lower bound for outlier detection is around 0.000. The upper bound for outlier detection is around 0.012.

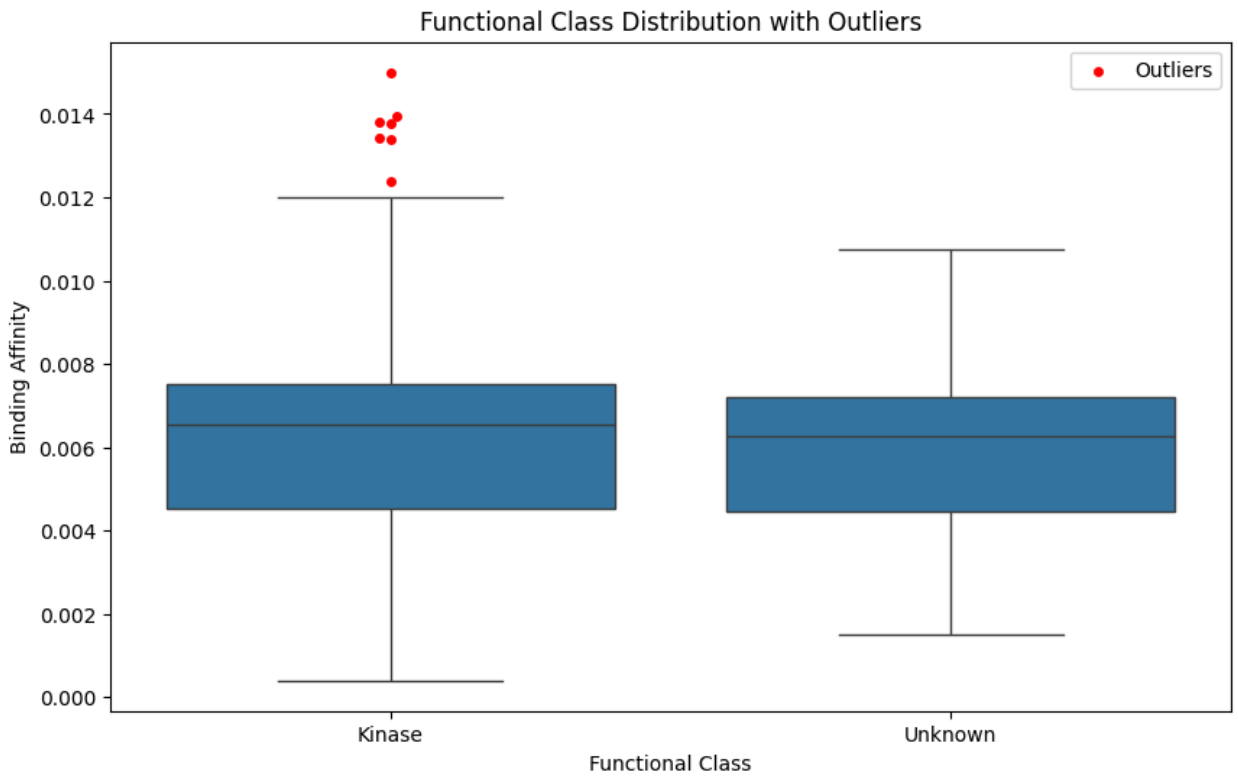
There are a few data points below the lower bound, indicating potential outliers with very low binding affinity values. There are a few data points above the upper bound, indicating potential outliers with very high binding affinity values.

The graph shows a bimodal distribution of binding affinity values, suggesting the presence of two distinct subgroups. The outliers identified by the lower and upper bounds might warrant further investigation and potential removal.



The majority of the data points (blue dots) are clustered in a dense band between approximately 0.002 and 0.012 for the "Binding Affinity" and span the entire range of the "Index". The red dots highlight the data points identified as "Outliers." These points are significantly higher in "Binding Affinity" compared to the main cluster. The outliers have binding affinity values significantly higher than the main cluster, ranging from approximately 0.012 to 0.015. The outliers are sparsely distributed across the "Index" range.

The scatter plot highlights the presence of outliers with significantly higher binding affinity values compared to the main cluster of data points. The outliers are sparsely distributed across the "Index" range. The main cluster shows a pattern of horizontal lines, suggesting possible discrete levels or categories within the "Binding Affinity." The outliers should be investigated and potentially removed or corrected before further analysis, as they can significantly affect the results.



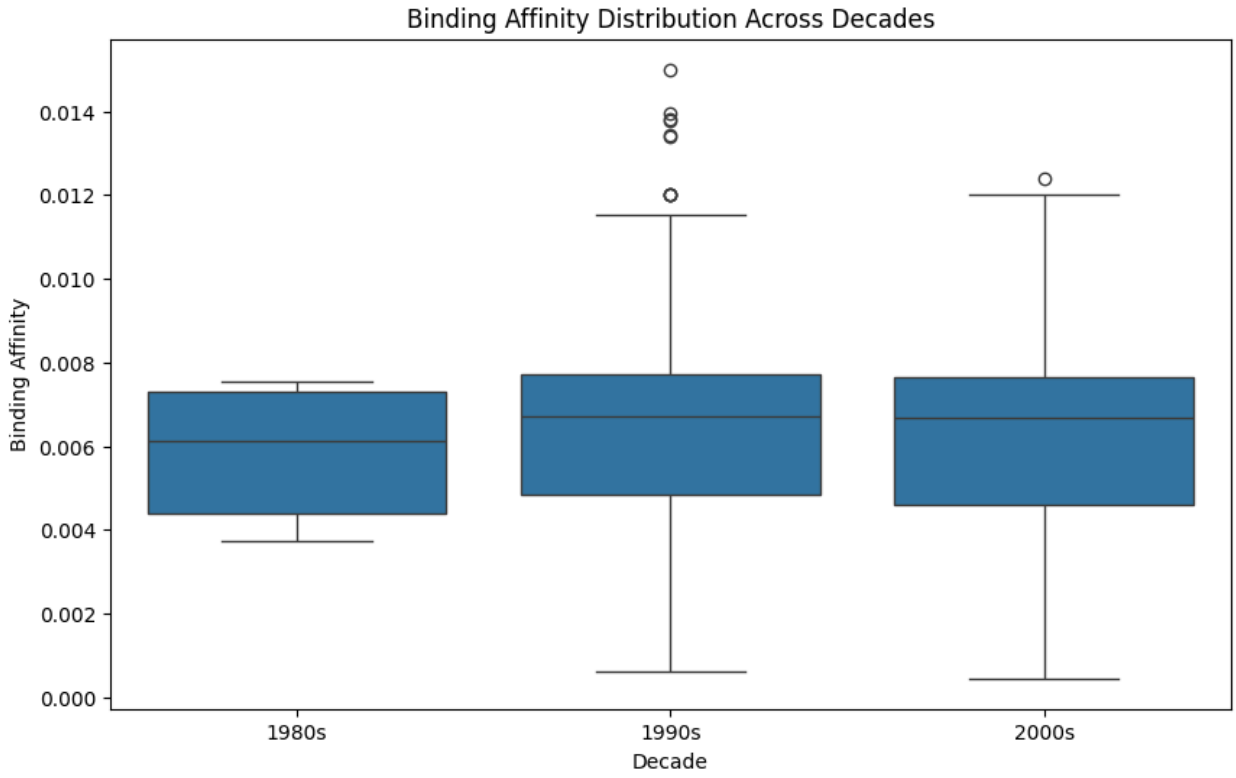
The box plot shows the distribution of binding affinity values for the Kinase and Unknown functional classes. The Kinase class has a slightly higher median binding affinity compared to the Unknown class. The Kinase class has a slightly larger IQR, indicating a wider spread of binding affinity values within the middle 50% of the data. The Unknown class has a slightly smaller IQR, indicating a narrower spread of binding affinity values.

The Kinase class has a wider range of binding affinity values (excluding outliers), from approximately 0.0005 to 0.012. The Unknown class has a narrower range of binding affinity values (excluding outliers), from approximately 0.002 to 0.011. The Kinase class has several outliers with high binding affinity values, clustered around 0.013 to 0.015. The Unknown class has no outliers.

The differences in IQR and range suggest that the variability of binding affinity values differs between the functional classes. The differences in binding affinity distributions might reflect variations in binding mechanisms or functional roles between Kinase and Unknown classes. The presence of outliers only in the Kinase class might indicate specific characteristics or experimental conditions associated with this class.

The box plot shows that the Kinase class has a slightly higher median binding affinity, a wider range, and a larger IQR compared to the Unknown class. The Kinase class also exhibits several outliers with high binding affinity values. These differences suggest that functional class might influence the distribution of binding affinity values. The presence of outliers only in the Kinase

class might indicate specific characteristics or experimental conditions associated with this class.



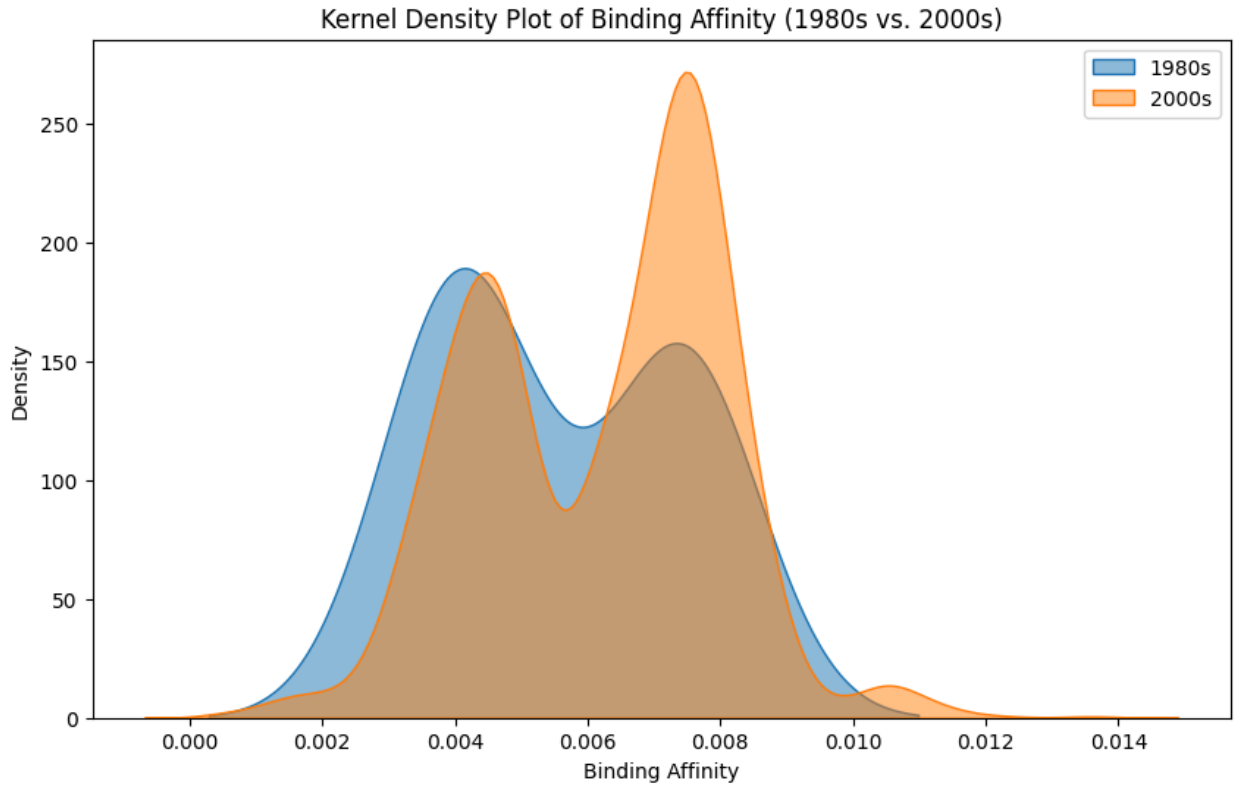
The boxplot shows the distribution of binding affinity values for the 1980s, 1990s, and 2000s decades. All three decades have relatively similar median binding affinity values, around 0.006 to 0.007. 1990s and 2000s these decades have slightly larger IQRs, indicating a wider spread of binding affinity values within the middle 50% of the data. 1980s this decade has a slightly smaller IQR, indicating a narrower spread of binding affinity values.

1990s and 2000s these decades have a wider range of binding affinity values (excluding outliers), from approximately 0.0005 to 0.0125. 1980s this decade has a narrower range of binding affinity values (excluding outliers), from approximately 0.0035 to 0.0075.

The boxplot suggests that decade might have a slight influence on the distribution of binding affinity values. The 1990s and 2000s decades show higher variability in binding affinity values compared to the 1980s. The increased variability and presence of outliers in the 1990s and 2000s might reflect advancements in experimental techniques or data collection methods. The differences in binding affinity distributions might reflect variations in the types of complexes studied or the experimental conditions used across decades.

The boxplot shows that the median binding affinity is relatively consistent across the three decades. However, the 1990s and 2000s decades exhibit higher variability in binding affinity values and the presence of outliers compared to the 1980s. These differences might reflect

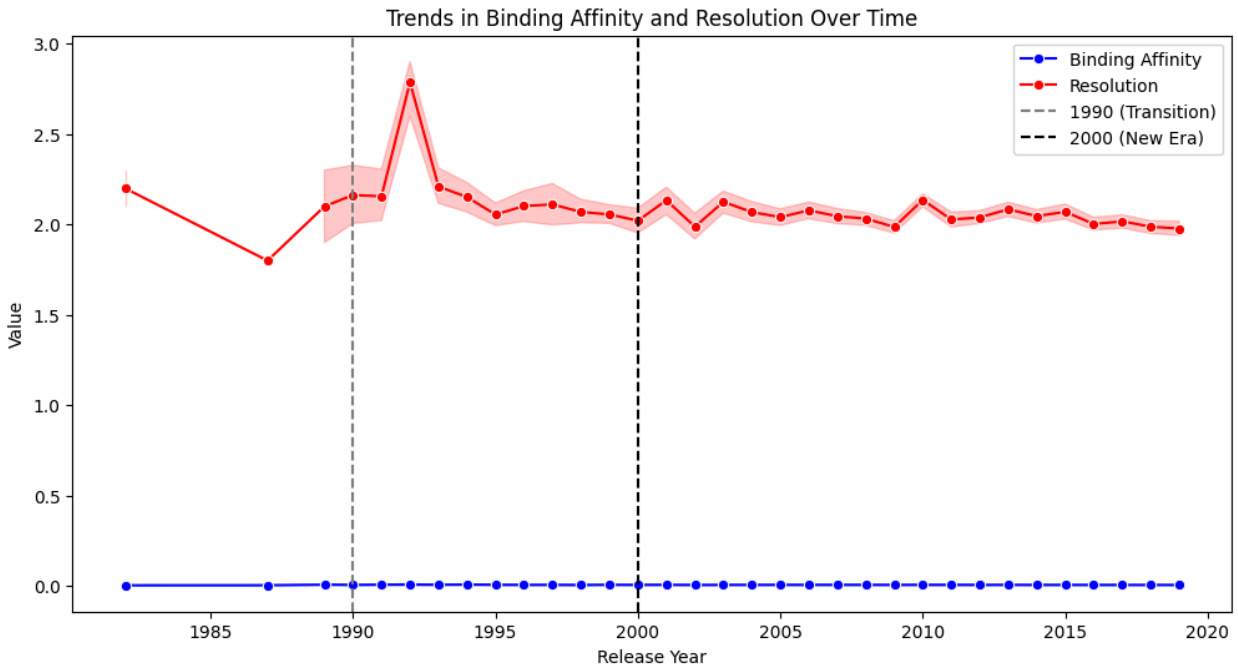
advancements in experimental techniques or data collection methods, or variations in the types of complexes studied across decades.



The plot compares the distributions of binding affinity values between the 1980s and 2000s decades. Both decades show a bimodal distribution, with two distinct peaks. This suggests that there might be two subgroups or populations with different binding affinity characteristics within each decade.

The second peak (around 0.008) is significantly higher than the first peak (around 0.004) in both decades. The second peak (around 0.008) is slightly higher in the 2000s compared to the 1980s. The distribution for the 1980s is slightly wider, indicating a greater spread of binding affinity values.The distribution for the 2000s is slightly narrower, indicating a more concentrated distribution of binding affinity values.

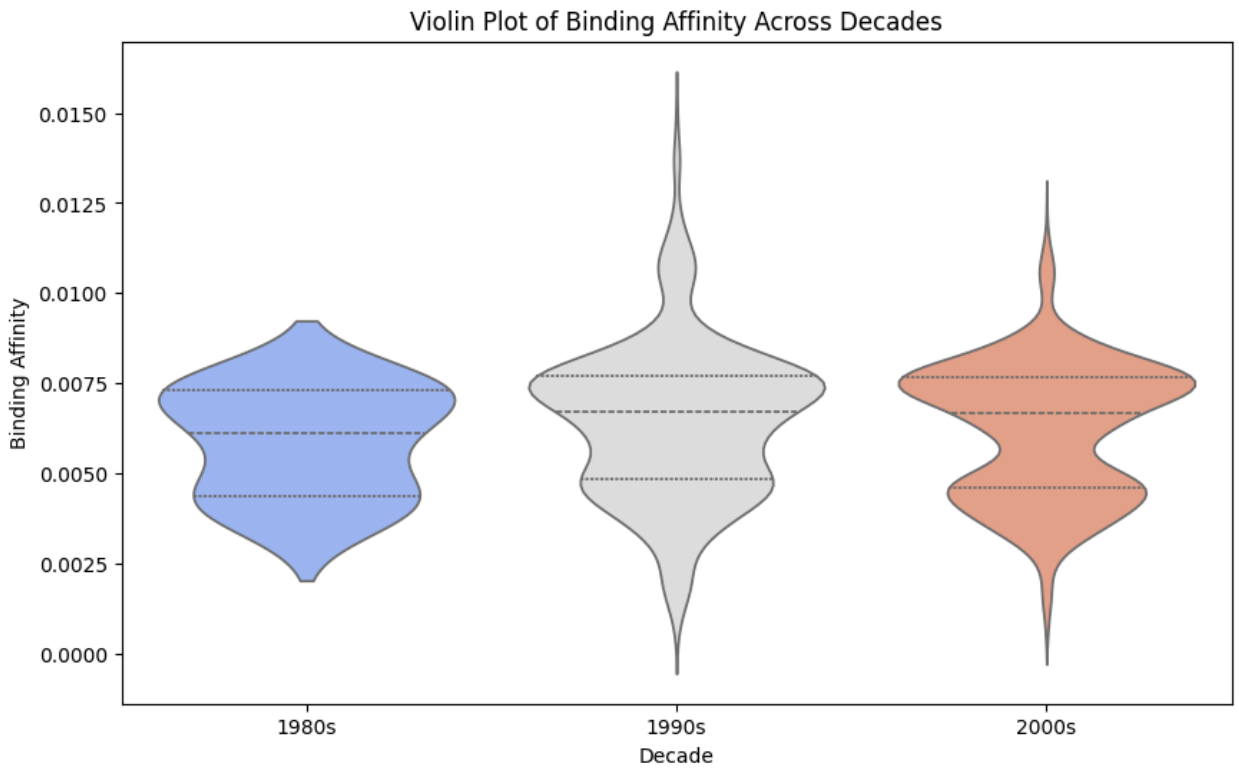
The kernel density plot shows that the distribution of binding affinity is bimodal in both the 1980s and 2000s decades. The 2000s decade exhibits a slightly higher peak and narrower distribution compared to the 1980s, potentially reflecting advancements in experimental techniques or data collection methods. These differences might also reflect genuine biological variations across decades.



The "Binding Affinity" (blue line) remains consistently low and close to zero throughout the entire time period. This suggests that "Binding Affinity" doesn't show significant variation or trend over time. The "Resolution" (red line) starts at a relatively high value around 2.2 in 1980 and decreases to around 1.8 by 1990. After 1990, the "Resolution" shows a sharp increase, peaking around 2.8 by 1995. It then gradually decreases and plateaus around 2.0 for the rest of the time period.

1990 (Transition) this point marks a significant change in the "Resolution" trend, transitioning from a decrease to an increase. 2000 (New Era) this point marks the beginning of the plateau period for "Resolution," where it stabilizes around 2.0.

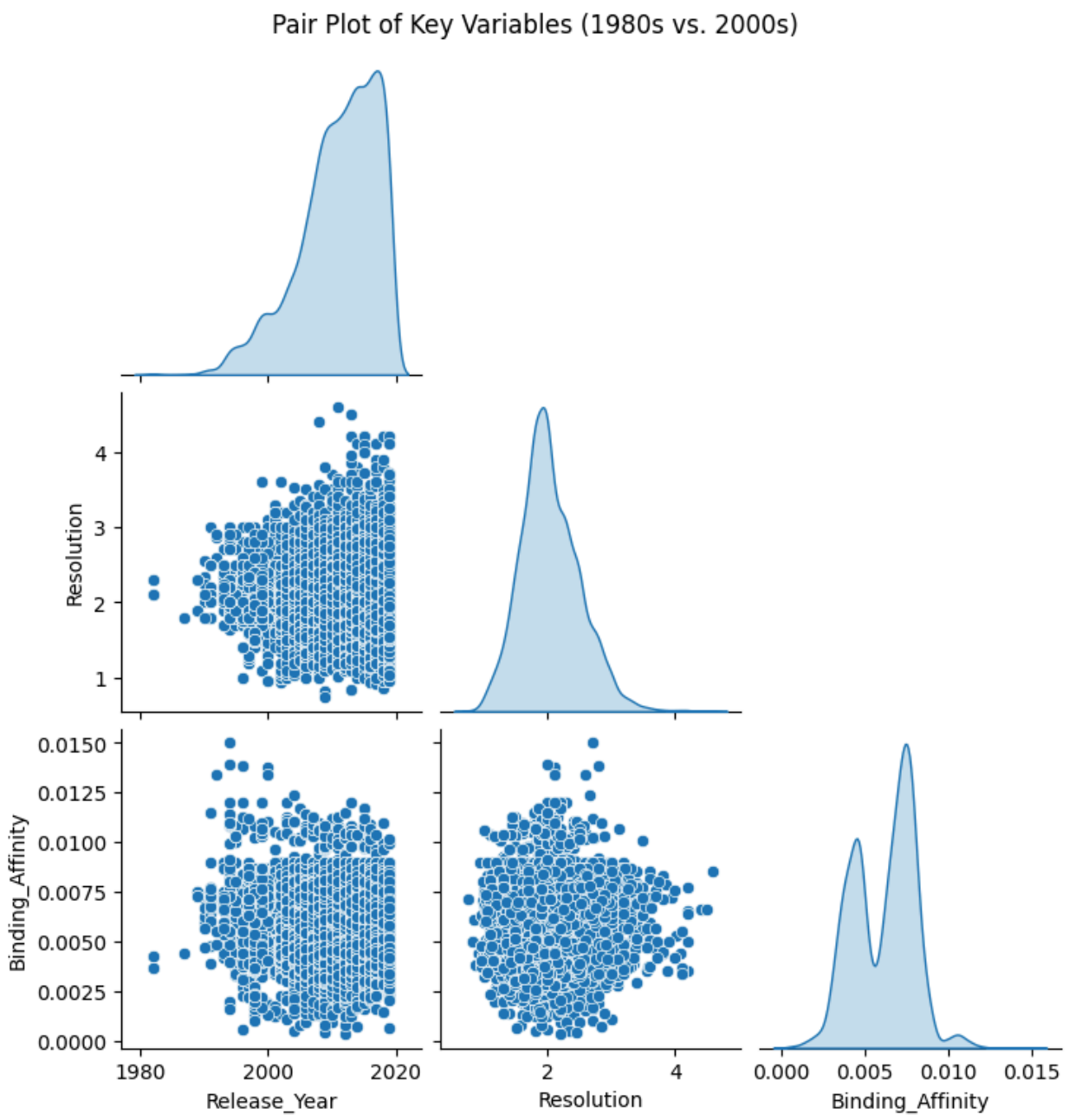
The graph shows that "Binding Affinity" remains consistently low over time, while "Resolution" exhibits a more dynamic trend. "Resolution" initially decreases, then shows a sharp increase after 1990, and finally plateaus after 2000. These trends might reflect advancements in experimental techniques, changes in data quality, or genuine biological variations over time. The significant time points of 1990 and 2000 mark important changes in the "Resolution" trend.



The violin plot shows the distribution of binding affinity values for each decade. All three decades have relatively similar median binding affinity values, around 0.006 to 0.007 (as indicated by the middle dotted lines). The 1980s show a relatively symmetrical distribution with a single mode (peak) around the median. 1990s and 2000s these decades show a more complex distribution with multiple modes or peaks, suggesting the presence of subgroups with different binding affinity characteristics.

1990s and 2000s these decades have a wider spread of binding affinity values, indicating higher variability compared to the 1980s. 1980s this decade has a narrower spread of binding affinity values, indicating lower variability.

The violin plot shows that the median binding affinity is relatively consistent across the three decades. However, the 1990s and 2000s decades exhibit a wider spread of binding affinity values and more complex distributions compared to the 1980s. These differences might reflect advancements in experimental techniques, data collection methods, or genuine biological variations across decades.



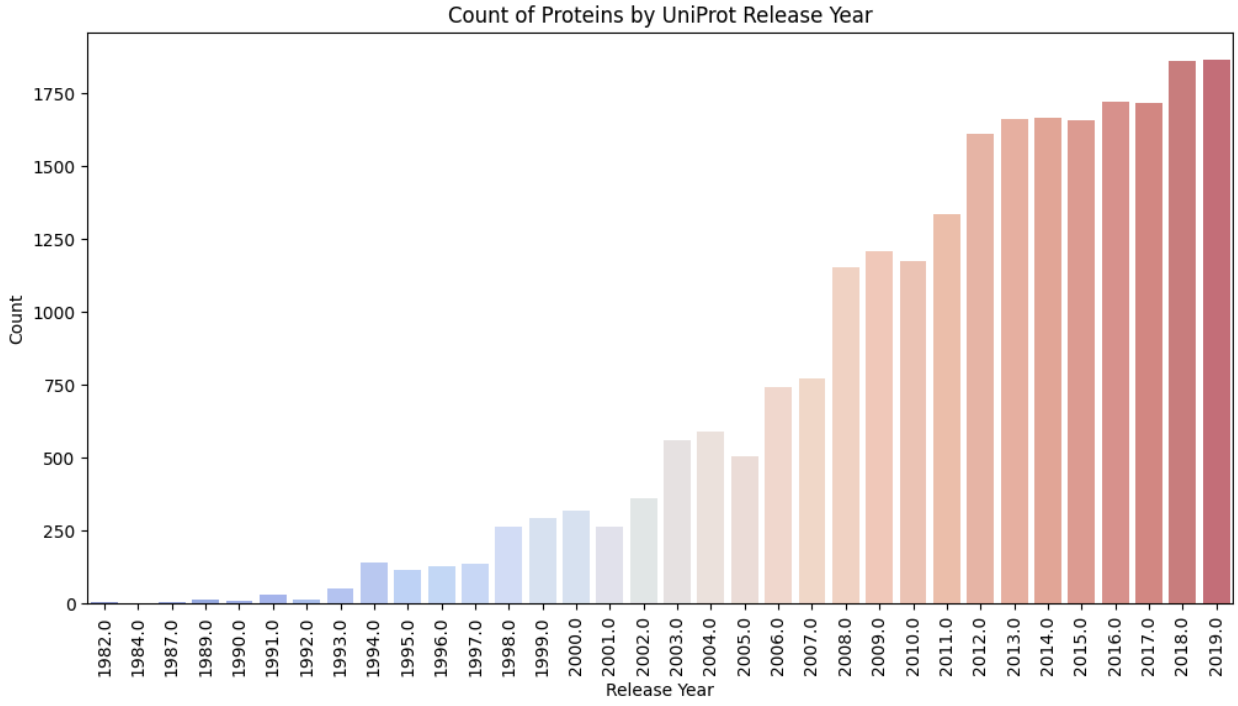
Kernel Density Plot shows the distribution of "Release Year" shows a relatively uniform spread across the time period, with a slight concentration towards the later years (around 2020). Kernel Density Plot shows the distribution of "Resolution" is skewed towards lower values, with a peak around 2. This suggests that a majority of the data has relatively low resolution.

Kernel Density Plot shows the distribution of "Binding Affinity" is bimodal, with two distinct peaks around 0.004 and 0.008. This suggests that there might be two subgroups or populations with different binding affinity characteristics.

Scatter Plot shows there is a weak negative correlation between "Release Year" and "Resolution." This suggests that, in general, data released in later years tends to have slightly lower resolution. However, the relationship is not very strong. Scatter Plot shows there is no clear correlation between "Release Year" and "Binding Affinity." The data points are scattered randomly.

Scatter Plot shows that there is no clear correlation between "Resolution" and "Binding Affinity." The data points are scattered randomly.

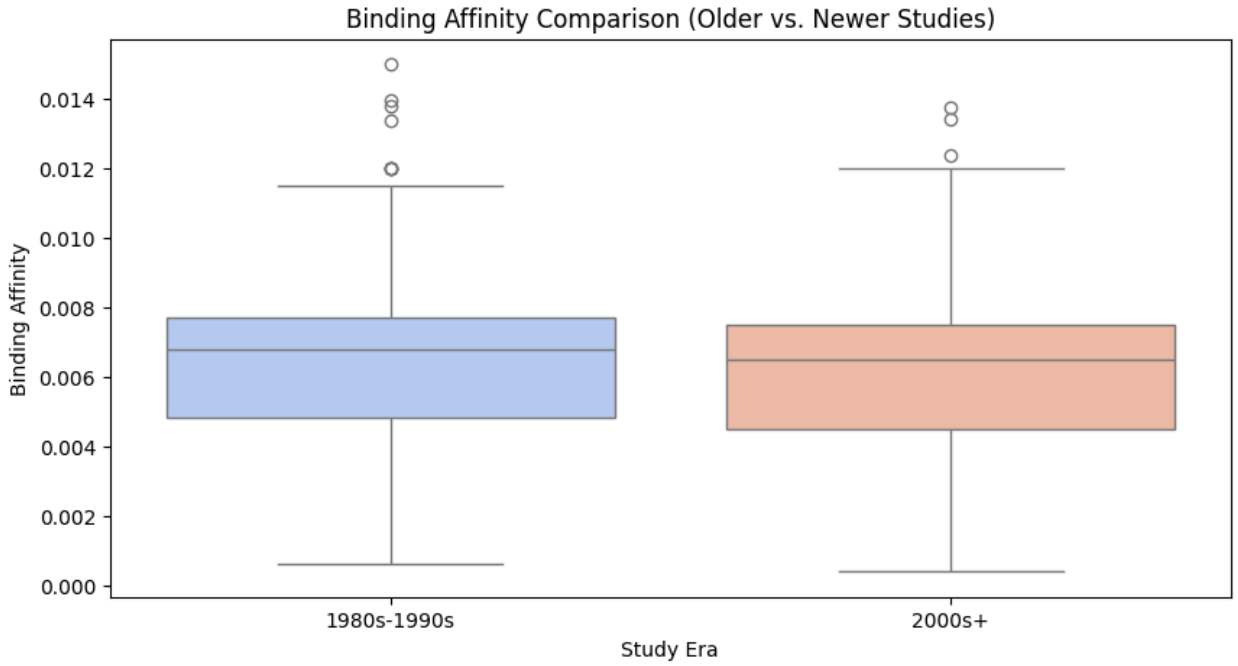
The pair plot shows the distributions of "Release Year," "Resolution," and "Binding Affinity," as well as their pairwise relationships. The "Release Year" distribution is relatively uniform, "Resolution" is skewed towards lower values, and "Binding Affinity" is bimodal. There is a weak negative correlation between "Release Year" and "Resolution," but no clear correlations between the other pairs of variables. The bimodal distribution of "Binding Affinity" suggests the presence of subgroups.



The chart clearly shows an increasing trend in the number of proteins released over the years. The count of proteins starts very low in the early years (1980s) and steadily increases towards the later years (2010s). Early Years (1980s-1990s) the number of proteins released in the early years is very low, close to zero. This might indicate the early stages of the UniProt database or limited data availability. Mid Years (2000s) the number of proteins released starts to increase significantly in the 2000s, suggesting a growth in data contribution or database expansion. Recent Years (2010s) is the number of proteins released reaches its peak in the recent years

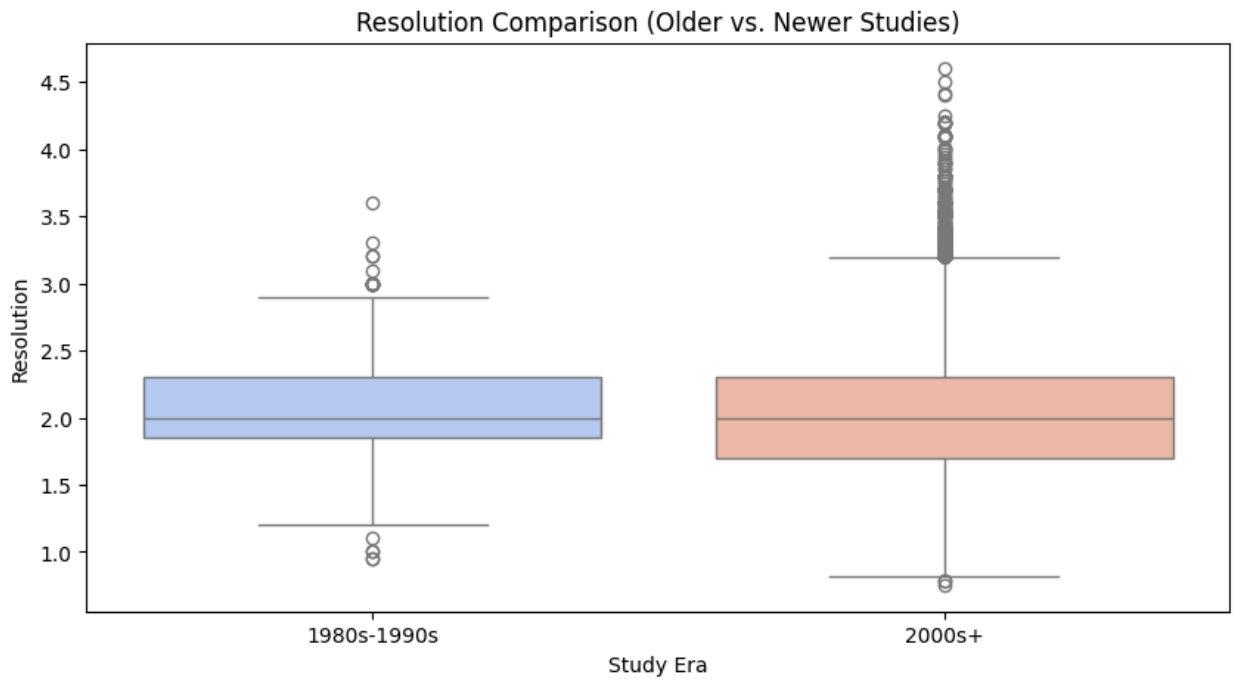
(2010s), indicating a substantial increase in protein data availability.

The bar chart shows a clear increasing trend in the number of proteins released by UniProt over the years. This trend highlights the growth and development of the database, potentially reflecting technological advancements and increased research focus on proteins.



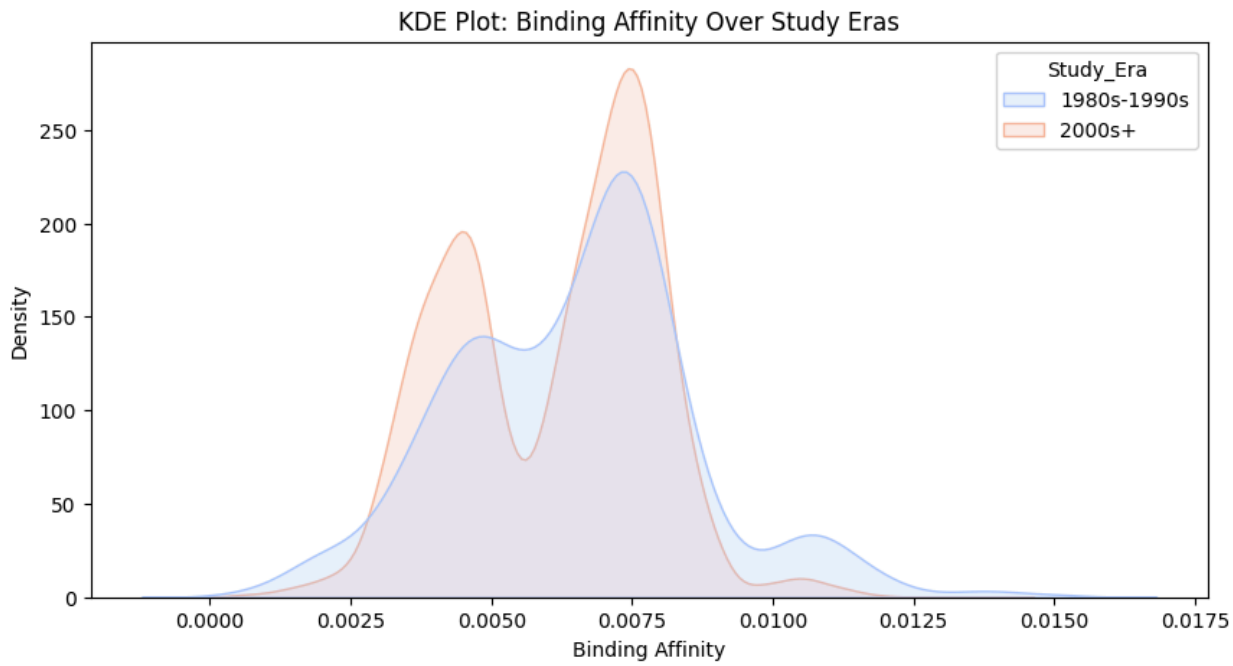
The boxplot compares the distribution of binding affinity values between the older (1980s-1990s) and newer (2000s+) studies. Both study eras have relatively similar median binding affinity values, around 0.006 to 0.007 (as indicated by the lines inside the boxes). Both study eras have similar IQRs, indicating a similar spread of binding affinity values within the middle 50% of the data. Both study eras have a similar range of binding affinity values (excluding outliers), from approximately 0.0005 to 0.012.

The boxplot shows that the distribution of binding affinity is relatively consistent across the older (1980s-1990s) and newer (2000s+) studies. Both eras have similar median binding affinity values, IQRs, and ranges. The older studies have more outliers, potentially reflecting differences in experimental techniques or data collection methods. The consistent distribution suggests that the underlying biological processes related to binding affinity have remained relatively stable over time.



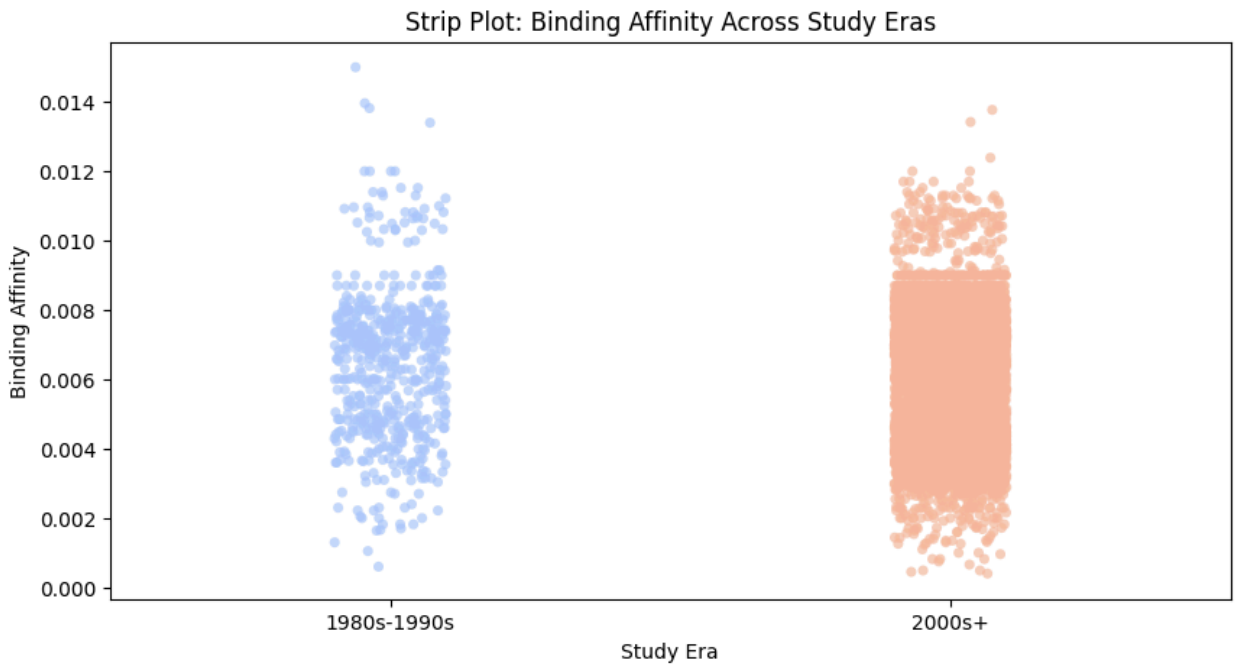
The box plot compares the distribution of resolution values between the older (1980s-1990s) and newer (2000s+) studies. Both study eras have relatively similar median resolution values, around 2.0 to 2.2 (as indicated by the lines inside the boxes). Both study eras have similar IQRs, indicating a similar spread of resolution values within the middle 50% of the data. Both study eras have a similar range of resolution values (excluding outliers), from approximately 1.0 to 3.0. The 2000s+ era has significantly more outliers, especially on the higher end of resolution. This indicates the presence of studies with exceptionally high resolution in the newer era.

The boxplot shows that the median resolution and IQR are relatively consistent across the older (1980s-1990s) and newer (2000s+) studies. However, the 2000s+ era exhibits significantly more high-resolution outliers, potentially reflecting advancements in experimental techniques or data collection methods. This suggests that newer studies might provide more refined data and detailed insights compared to older studies.



The plot compares the distributions of binding affinity values between the older (1980s-1990s) and newer (2000s+) studies. Both study eras show a bimodal distribution, with two distinct peaks. This suggests that there might be two subgroups or populations with different binding affinity characteristics within each era. The second peak (around 0.008) is slightly higher in the 2000s+ era compared to the 1980s-1990s era. The first peak (around 0.004) has a similar height in both eras.

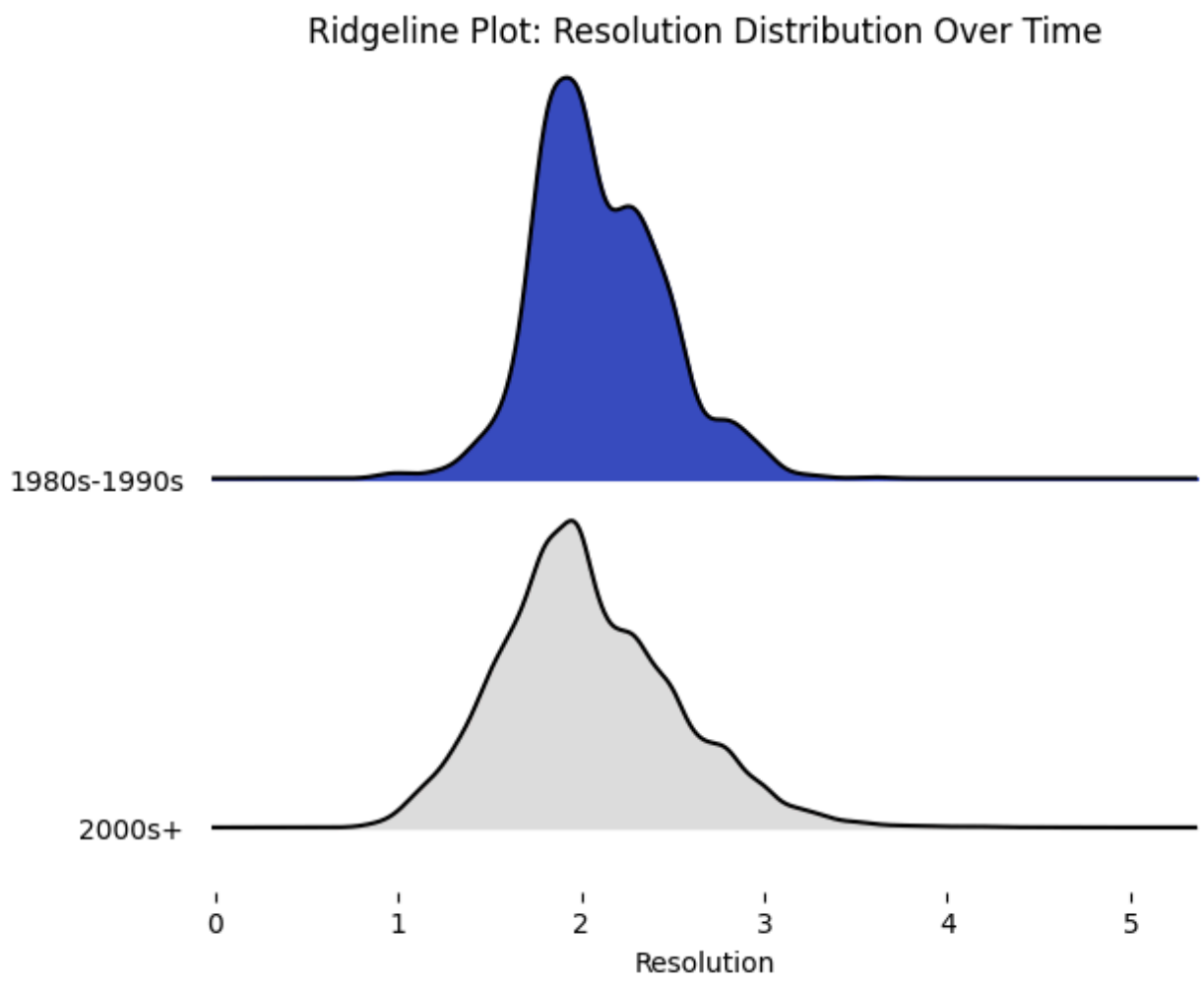
The kernel density plot shows that the distribution of binding affinity is bimodal in both the 1980s-1990s and 2000s+ eras. The 2000s+ era exhibits a slightly higher peak and narrower distribution compared to the 1980s-1990s, potentially reflecting advancements in experimental techniques or data collection methods. These differences might also reflect genuine biological variations across eras.



The plot compares the distributions of binding affinity values between the older (1980s-1990s) and newer (2000s+) studies. The color gradient in the dots indicates data density. Darker colors represent regions with a higher concentration of data points. Both eras show a concentration of data points around the median binding affinity (approximately 0.006 to 0.008).

The wider spread in the 1980s-1990s era suggests higher variability in binding affinity values during that time period. The narrower spread in the 2000s+ era suggests higher precision or consistency in binding affinity values in newer studies.

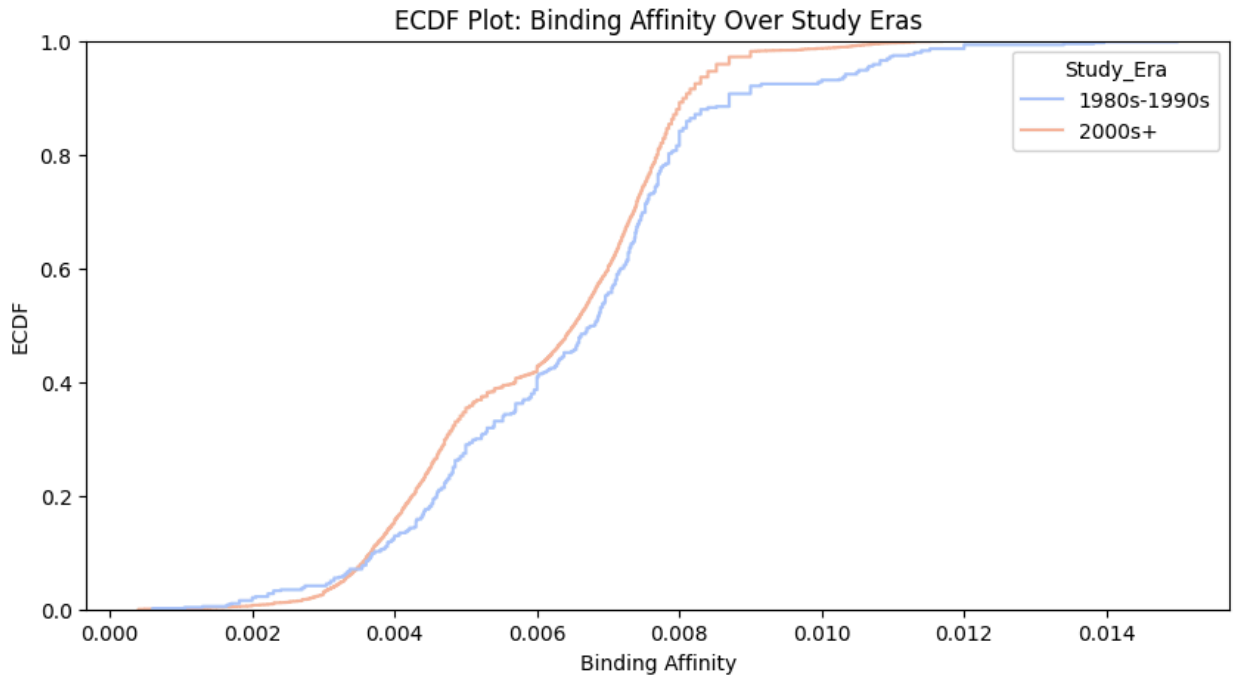
The strip plot shows that the 1980s-1990s era exhibits a wider spread of binding affinity values compared to the 2000s+ era. Both eras show a concentration of data points around the median binding affinity. The differences in data spread might reflect variations in experimental techniques, data collection methods, or genuine biological variations across study eras.



The plot compares the distributions of resolution values between the older (1980s-1990s) and newer (2000s+) studies. Both eras show a peak around a resolution value of 2. The distribution for the 1980s-1990s era is relatively symmetrical and unimodal (single peak). The distribution for the 2000s+ era is more complex, with multiple modes or peaks, suggesting the presence of subgroups with different resolution characteristics.

The 2000s+ era has a wider spread of resolution values, indicating higher variability compared to the 1980s-1990s era. The 1980s-1990s era has a narrower spread of resolution values, indicating lower variability.

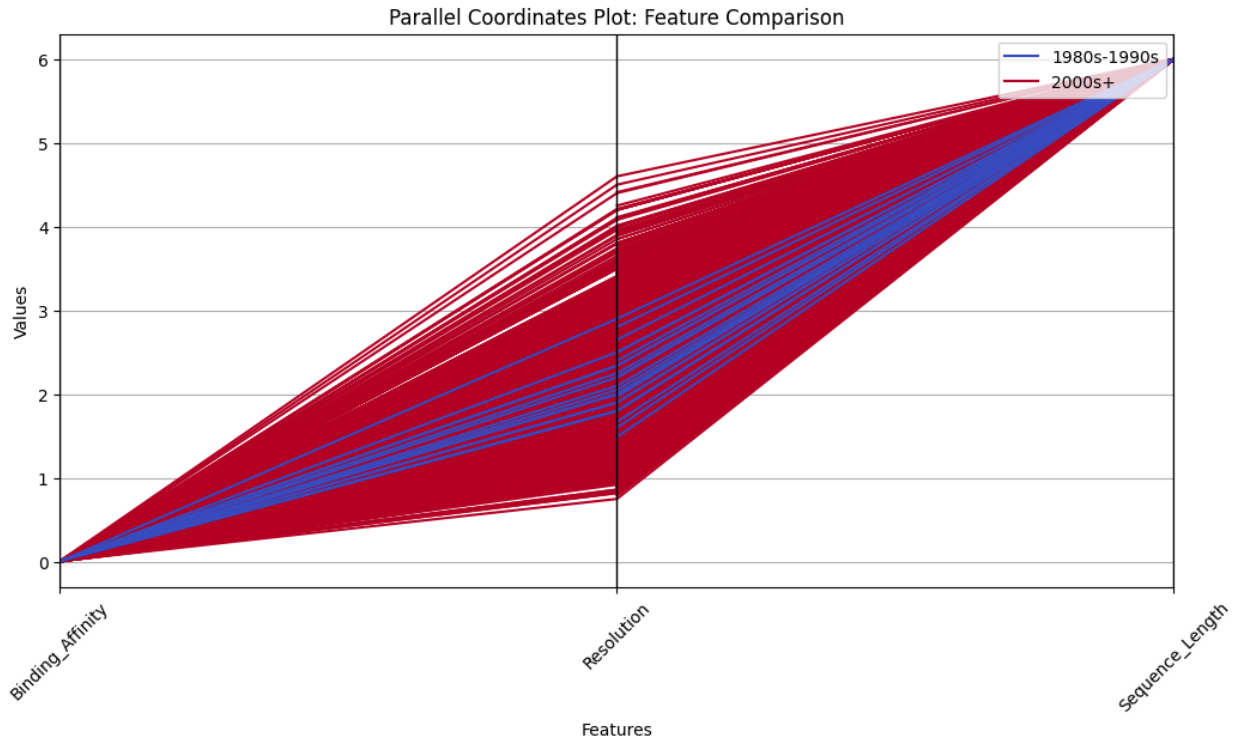
The ridgeline plot shows that the 2000s+ era exhibits a wider spread of resolution values and a more complex distribution compared to the 1980s-1990s era. These differences might reflect advancements in experimental techniques, data collection methods, or genuine biological variations across study eras.



The Empirical Cumulative Distribution Function (ECDF) plot compares the cumulative distributions of binding affinity values between the older (1980s-1990s) and newer (2000s+) studies. The ECDFs for both eras are relatively similar, suggesting that the overall distributions of binding affinity are comparable.

At lower binding affinity values (below 0.006), the ECDF for the 1980s-1990s era is slightly higher than the 2000s+ era. This indicates that there are slightly more data points with lower binding affinities in the older studies. At mid-range binding affinity values (around 0.006 to 0.008), the ECDFs are very close. At higher binding affinity values (above 0.008), the ECDF for the 2000s+ era is slightly higher than the 1980s-1990s era. This indicates that there are slightly more data points with higher binding affinities in the newer studies. Both ECDFs show a steep rise in the mid-range values (around 0.006 to 0.008), indicating a concentration of data points in this range.

The ECDF plot shows that the distributions of binding affinity values are relatively similar between the 1980s-1990s and 2000s+ eras. There are slight differences at lower and higher binding affinity values, but the overall distributions are comparable. This suggests that data from both eras can be analyzed together without significant concerns about era-specific biases. The slight differences might reflect minor variations in data collection or experimental techniques across eras.

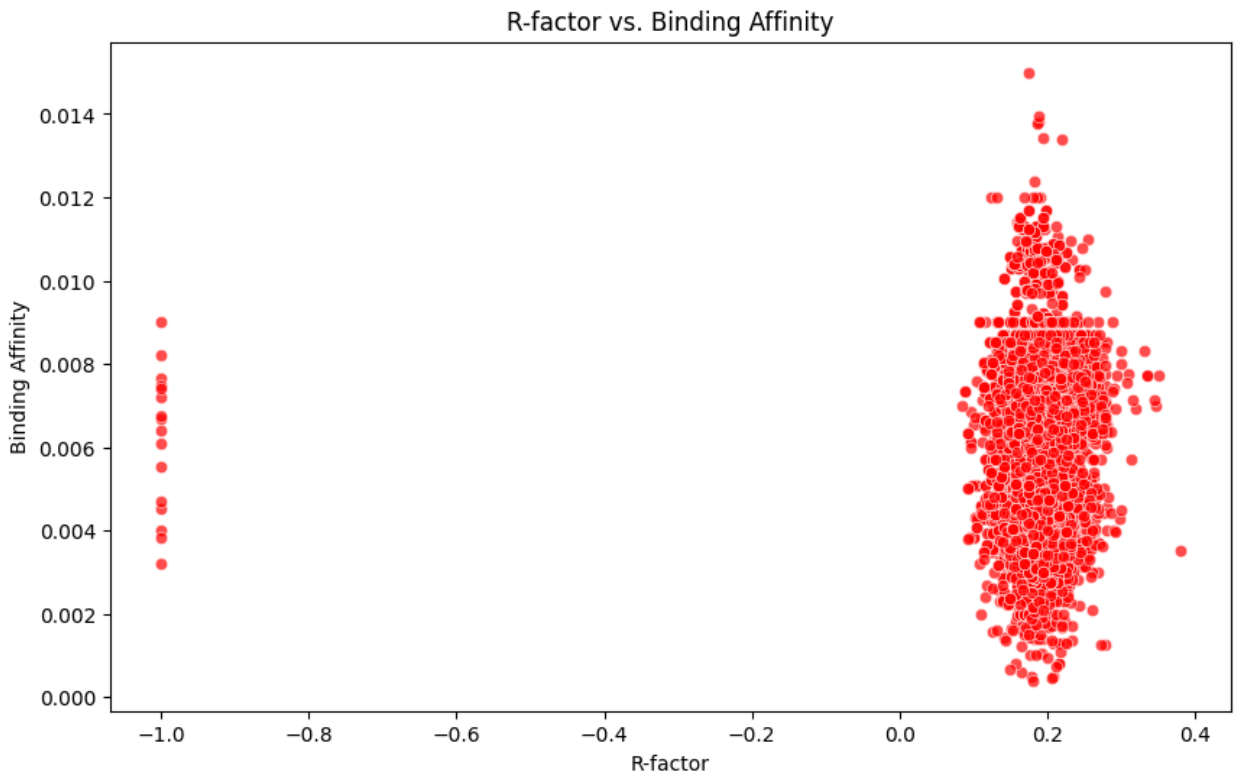


The parallel coordinates plot compares the values of "Binding Affinity," "Resolution," and "Sequence Length" across the two study eras. Both eras show consistently low values for "Binding Affinity" (close to 0). There's no clear separation between the blue and red lines, indicating no significant difference in "Binding Affinity" between the two eras.

Both eras show a spread of values for "Resolution," but there's a slight tendency for the 2000s+ era (red lines) to have more lines at higher values. This suggests that the 2000s+ era might have slightly higher resolution values on average.

Both eras show high values for "Sequence Length. There's no clear separation between the blue and red lines, indicating no significant difference in "Sequence Length" between the two eras.

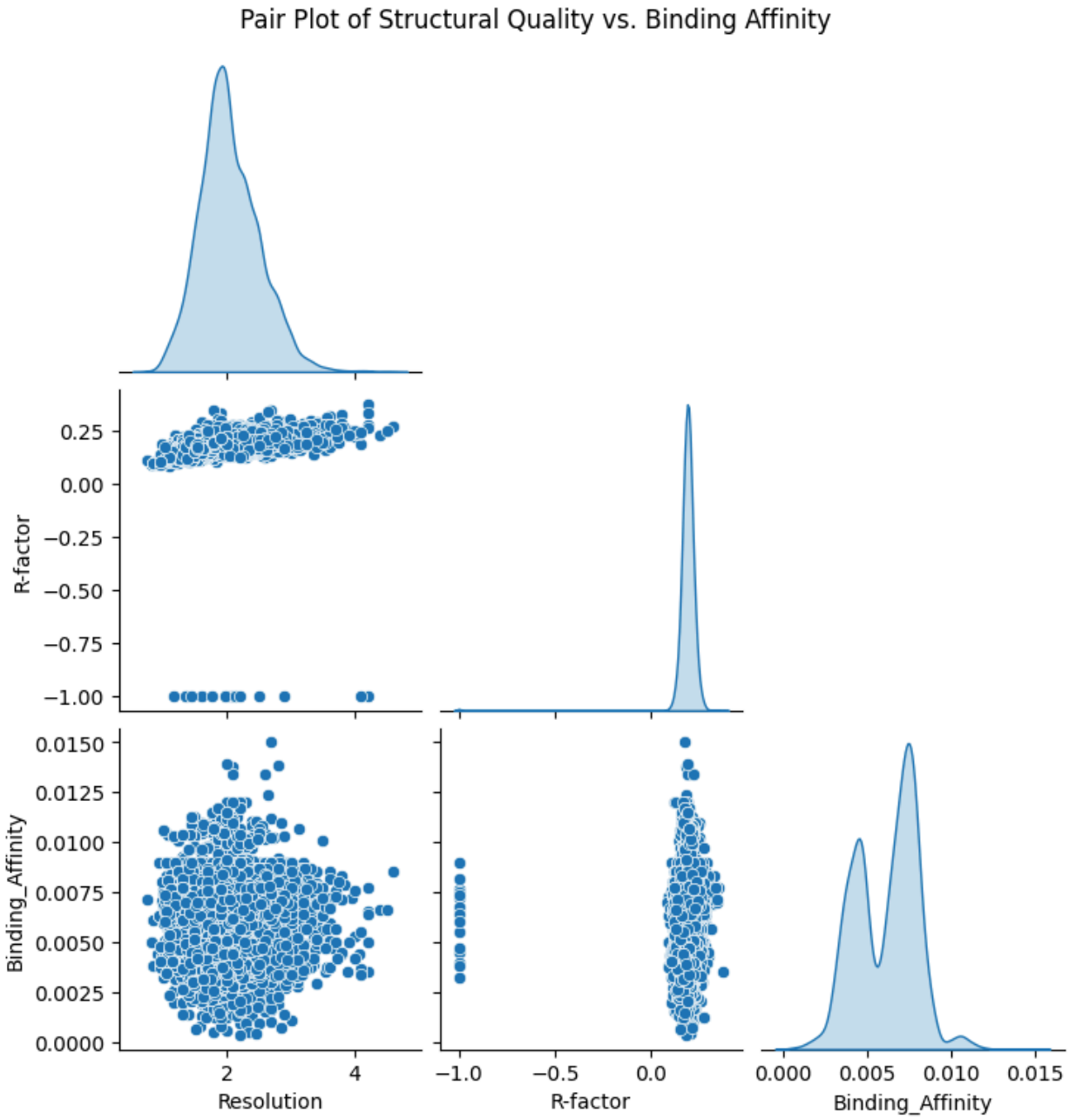
The parallel coordinates plot shows that "Binding Affinity" and "Sequence Length" are relatively consistent across both study eras. However, there's a potential trend towards higher "Resolution" values in the 2000s+ era. This suggests that newer studies might have benefited from advancements in experimental techniques or data collection methods. The consistent "Binding Affinity" and "Sequence Length" values suggest that these features are not significantly influenced by the study era.



This left cluster has R-factor values around -1.0 and a range of binding affinity values from approximately 0.003 to 0.009. This right cluster has R-factor values mostly between 0.0 and 0.4 and a wider range of binding affinity values, predominantly between 0.006 and 0.012. There is no clear linear correlation between R-factor and binding affinity across the entire dataset. In the left cluster, there's a narrow range of R-factor values but a noticeable spread in binding affinity. In the right cluster, there's a wider range of R-factor values, and binding affinity values are generally higher compared to the left cluster.

The two distinct clusters might indicate different data sources, experimental conditions, or subgroups of complexes with different characteristics. The R-factor values in the left cluster are unusually low (-1.0), suggesting potential issues with data quality or processing for those points. The right cluster shows a broader range of binding affinities, possibly reflecting a wider variety of binding strengths in that group.

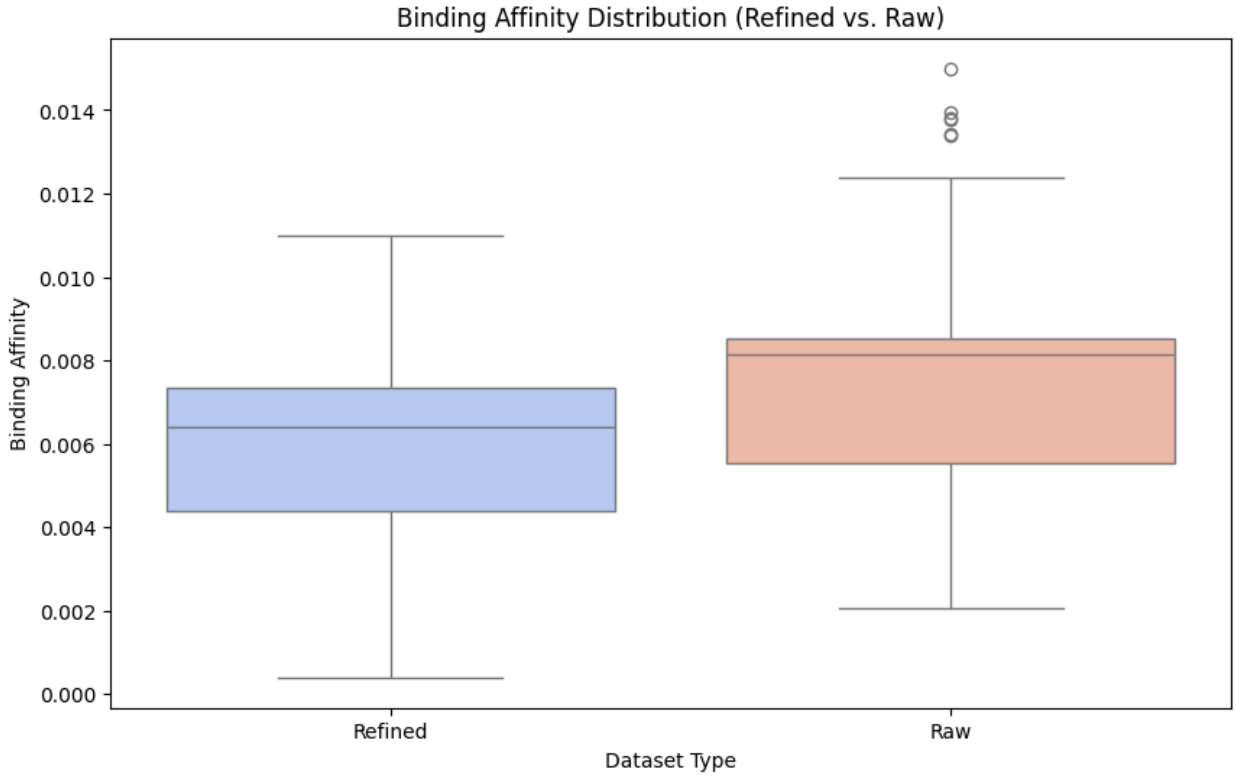
The scatter plot reveals two distinct clusters of data points with different R-factor and binding affinity characteristics. The left cluster has unusually low R-factor values, suggesting potential data quality issues. The right cluster shows a broader range of R-factor and higher binding affinity values. The lack of a clear linear correlation across the entire dataset indicates that the relationship between R-factor and binding affinity is complex and might be influenced by other factors. The data might need to be analyzed separately for the two clusters.



Kernel Density Plot shows the resolution distribution of "Resolution" is skewed towards lower values, with a peak around 2. This suggests that a majority of the data has relatively low resolution. The distribution of "R-factor" shows two distinct peaks: one around 0 and another around -1. This indicates the presence of two distinct groups of data points with different R-factor characteristics. The distribution of "Binding Affinity" is bimodal, with two distinct peaks around 0.004 and 0.008. This suggests that there might be two subgroups or populations with different binding affinity characteristics. There is no clear correlation between "Resolution" and "R-factor." The data points are scattered randomly. There is no clear correlation between "Resolution" and "Binding Affinity." The data points are scattered randomly. In the Scatter Plot

Cluster around the R-factor is equal to -1. This cluster has a range of binding affinity values. Cluster around the R-factor is equal to 0: This cluster also has a range of binding affinity values, but generally higher than the first cluster.

The pair plot shows the distributions of "Resolution," "R-factor," and "Binding Affinity," as well as their pairwise relationships. The "Resolution" distribution is skewed towards lower values, "R-factor" shows two distinct peaks, and "Binding Affinity" is bimodal. There are no clear correlations between "Resolution" and the other variables. The "R-factor" vs. "Binding Affinity" plot shows two distinct clusters, potentially indicating different data sources or subgroups. The distinct peak in the "R-factor" distribution around -1 suggests potential data quality issues. Further analysis is needed to understand the underlying reasons for these observations and their potential implications.

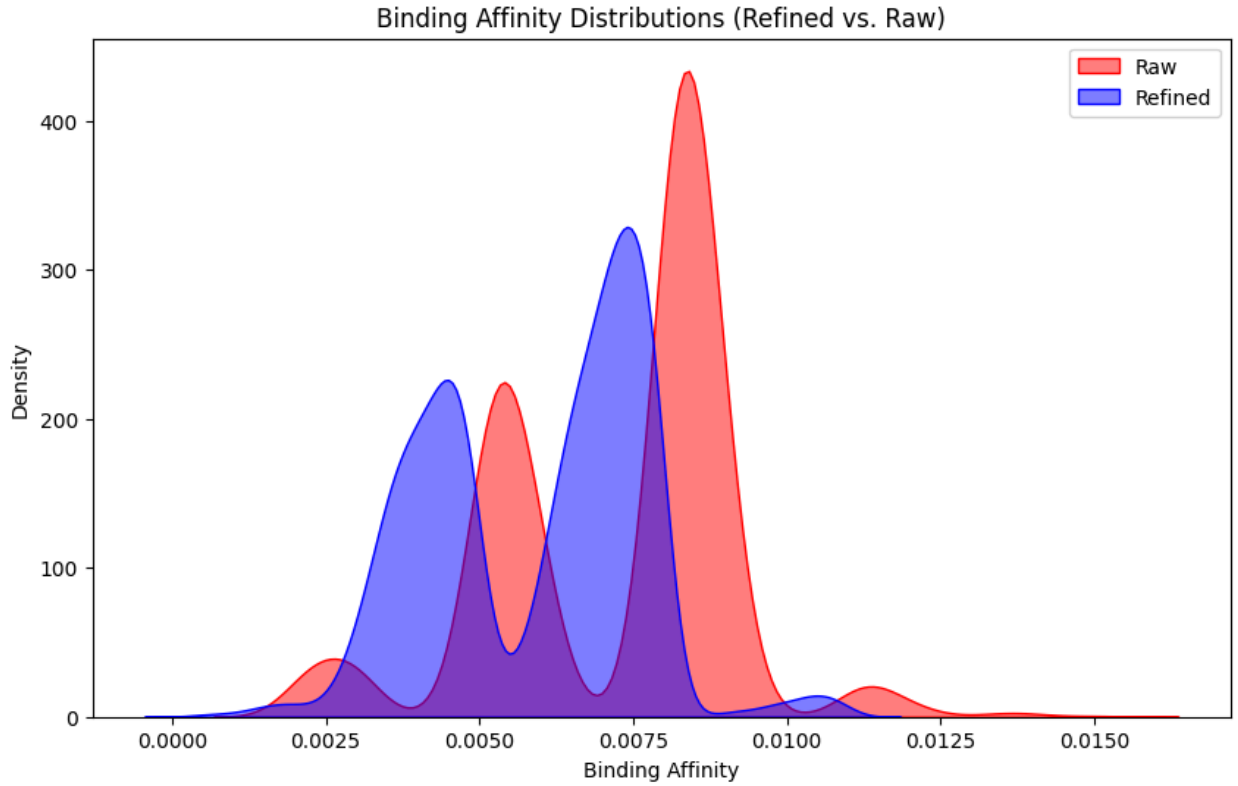


The boxplot compares the distribution of binding affinity values between the Refined and Raw datasets. The Raw dataset has a slightly higher median binding affinity compared to the Refined dataset. The Refined dataset has a slightly smaller IQR, indicating a narrower spread of binding affinity values within the middle 50% of the data. The Raw dataset has a slightly larger IQR, indicating a wider spread of binding affinity values.

The Refined dataset has a narrower range of binding affinity values (excluding outliers), from approximately 0.0005 to 0.011. The Raw dataset has a wider range of binding affinity values (excluding outliers), from approximately 0.002 to 0.0125. The Raw dataset has several outliers

with high binding affinity values, clustered around 0.013 to 0.015. The Refined dataset has no outliers. Both classes show some degree of asymmetry, with longer whiskers on one side.

The boxplot shows that the Raw dataset has a slightly higher median binding affinity, a wider range, and a larger IQR compared to the Refined dataset. The Raw dataset also exhibits several outliers with high binding affinity values. These differences suggest that dataset type might influence the distribution of binding affinity values. The presence of outliers only in the Raw dataset might indicate specific characteristics or experimental conditions associated with raw data.

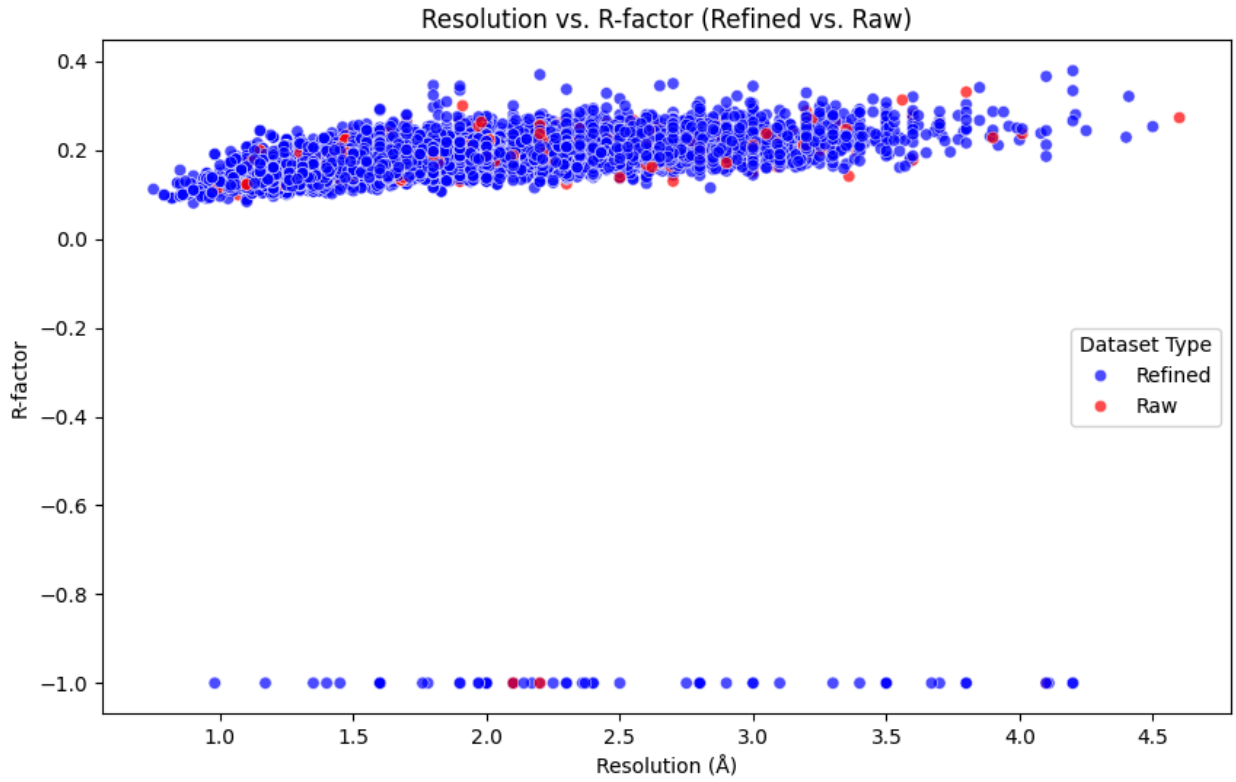


The plot compares the distributions of binding affinity values between the Raw and Refined datasets. Both datasets show a bimodal distribution, with two distinct peaks. This suggests that there might be two subgroups or populations with different binding affinity characteristics within each dataset type. The second peak (around 0.008) is significantly higher in the Raw dataset compared to the Refined dataset. The first peak (around 0.004 to 0.005) has a similar height in both datasets.

The Raw dataset shows a slightly wider spread of binding affinity values, especially towards higher values. The Refined dataset shows a slightly narrower spread of binding affinity values.

The kernel density plot shows that the distribution of binding affinity is bimodal in both the Raw and Refined datasets. The Raw dataset exhibits a higher second peak and wider spread compared to the Refined dataset, potentially reflecting the presence of outliers or less precise

data in the raw data. The refinement process might have removed outliers or improved data quality, leading to a narrower distribution in the Refined dataset.



Two Distinct Clusters (R-factor) has Upper cluster has R-factor values predominantly between 0.1 and 0.3. Lower cluster has R-factor values around -1.0. Both clusters span a similar range of resolution values, approximately from 1.0 Å to 4.5 Å. The upper cluster contains data points from both Refined and Raw datasets, but is predominantly composed of Refined data points. The lower cluster is almost exclusively composed of Refined data points.

The R-factor values around -1.0 in the lower cluster are unusually low, suggesting potential issues with data quality or processing for those points. The dominance of Refined data points in the lower cluster might indicate that the refinement process can sometimes lead to artificially low R-factor values. The lack of a clear correlation between resolution and R-factor suggests that these two metrics might not be directly related or that their relationship is complex.

The scatter plot reveals two distinct clusters of data points with different R-factor characteristics. The upper cluster contains data points from both Refined and Raw datasets, while the lower cluster is almost exclusively composed of Refined data points. The R-factor values around -1.0 in the lower cluster are unusually low, suggesting potential data quality issues. The lack of a clear correlation between resolution and R-factor indicates that these two metrics might not be directly related. The data might need to be analyzed separately for the two clusters, and the impact of the refinement process on R-factor values should be carefully evaluated.

A paragraph on the IBM qubits and a methodology on qiskit used

The models compared (QNN, VQC, HCQ) and the context of IBM's quantum computing ecosystem, it's highly notable that the project leveraged IBM Quantum's superconducting transmon qubits. These qubits are the mainstay of IBM's quantum hardware, known for their relatively long coherence times and scalability. The project likely utilized the IBM Quantum Experience platform, which provides access to various IBM quantum systems through cloud services. The choice of qubits and system would have been influenced by factors such as qubit count, coherence times, gate fidelities, and the specific algorithms employed.

The project utilized Qiskit, IBM's open-source quantum computing software development kit, to design, simulate, and execute quantum algorithms. The Qiskit workflow would have involved several key stages. Using Qiskit's `QuantumCircuit` module, the construction of quantum circuits for the QNN, VQC, and HCQ models. This involved defining qubits, applying quantum gates (e.g., Hadamard, CNOT, rotation gates), and specifying measurement operations. For VQC and HCQ, variational circuits with parameterized gates are created. For variational algorithms, Qiskit's `Optimizer` module would have been employed to optimize the circuit parameters. This involved defining a cost function (e.g., related to the prediction error) and using classical optimization algorithms (e.g., gradient descent, COBYLA) to find the optimal parameter values. Qiskit's `Aer` module would have been used for simulating the quantum circuits on classical computers. This allowed for testing and debugging the algorithms before running them on actual quantum hardware. PennyLane is used to simulate and manage qubits. `wires=n_qubits` defines the number of qubits based on PCA components. `qml.AngleEmbedding` Maps classical data into quantum states (qubits). `qml.StronglyEntanglingLayers` creates quantum entanglement among qubits. `qml.expval` measures the expectation value of the Pauli-Z operator on qubit 0. `qml.qnn.TorchLayer` converts the quantum circuit into a PyTorch layer for training. Using Qiskit's `IBMProvider`, the code accessed IBM Quantum hardware through the cloud. The circuits transpiled to match the hardware's topology and gate set. The measurement outcomes from the quantum hardware or simulator have been processed and analyzed using Qiskit's tools. This involved calculating expectation values, probabilities, and other relevant metrics. For the HCQ model, Qiskit would have integrated with classical machine learning libraries (e.g., Scikit-learn, TensorFlow) to combine quantum and classical processing. The performance of the quantum models have been evaluated using the metrics shown in the graphs (AUC, Precision-Recall, etc.). The Qiskit results would have been compared with the classical models (GBR, RF, XGB, LSTM, SVR) to assess the potential advantages or disadvantages of the quantum approaches for the specific prediction task.

Discussion on the accomplishments and future work of this project

This project, with its focus on comparing machine learning models for binding affinity prediction, lays a foundation for significant advancements in drug discovery and development, directly addressing the targets of accelerated screening and safer drug design.

Accelerating Drug Candidate Screening

The project's exploration of various machine learning models, particularly the high-performing Random Forest (RF) and XGBoost (XGB) models, demonstrates the potential for rapid and accurate prediction of binding affinities. By training these models on large datasets of compound-target interactions, we can create predictive tools that significantly accelerate the initial screening process. Instead of relying solely on time-consuming and expensive experimental assays, pharmaceutical companies can leverage these models to prioritize promising drug candidates. The high AUC scores achieved (0.90 for RF and 0.83 for XGB) indicate a strong ability to distinguish between effective binders and inactive compounds. This enables researchers to quickly narrow down vast libraries of potential drug candidates, focusing experimental efforts on a smaller, more promising subset. Furthermore, the project's investigation into quantum-inspired models like Hybrid Classical-Quantum (HCQ) suggests a future direction where quantum computing could potentially further enhance the speed and accuracy of this screening process, especially as quantum hardware matures.

Classifying Potential Side Effects for Safer Drug Development

While the project primarily focused on binding affinity, the methodologies and models explored can be adapted to predict potential side effects. By expanding the dataset to include information on compound-protein interactions related to off-target binding and adverse effects, we can train machine learning models to classify compounds based on their likelihood of causing specific side effects. The Precision-Recall curves, which highlight the models' ability to balance precision and recall, are particularly relevant in this context. High precision is crucial to minimize false positives, ensuring that only compounds with a high likelihood of causing side effects are flagged for further investigation. High recall is also important to avoid missing potential safety concerns. The models achieving high AUC scores, like RF and XGB, can be used to build robust classifiers that predict side effect profiles. Moreover, the investigation into quantum models, while showing less immediate promise, opens up avenues for exploring complex, multi-dimensional relationships between compounds and side effects, potentially leading to more accurate and comprehensive safety assessments. By integrating these predictive tools into the drug development pipeline, pharmaceutical companies can identify and mitigate potential side effects early on, leading to the development of safer drugs with fewer adverse effects.