**HackBio Stage 3**

**Computational Analysis of Chemical Space: A PCA, Clustering, and Docking Score Prediction Approach**

Tyler Maire1, Mahpara Abid2, Peter Imonte3, Gracious Isoah4, Terry Ngala5, Mikaya Sheild6

|  |
| --- |
| **Abstract**  Understanding the chemical properties that influence molecular binding to protein targets is crucial for drug discovery. This study explores a dataset of over 10,000 compounds docked against adenosine deaminase (ADA), an enzyme linked to multiple diseases. We employ Principal Component Analysis (PCA) and K-means clustering to map the chemical space of these compounds and identify clusters with strong binding affinities. Additionally, we use machine learning regression models to predict docking scores, highlighting key molecular descriptors that influence binding. Our findings reveal distinct chemical clusters and the molecular properties that contribute to strong ADA binding, paving the way for more efficient drug screening strategies.  **Keywords**  Molecular Docking — Principal Component Analysis — K-Means Clustering — Random Forest — Drug Discovery |
| [[1]](#footnote-1)*University of Maryland Global Campus , College Park, Maryland, United States*  *2Quaid-i-Azam University, Islamabad, Pakistan*  *3University of Port Harcourt, Rivers State, Nigeria*  *4University of Benin, Edo State, Nigeria*  *5The Technical University of Kenya, Nairobi, Kenya*  *6Arizonia State University, Phoenix, Arizonia, United States*  \***Corresponding author**: tyler.maire1@gmail.com |

#### Contents

1. **Introduction** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **1**

2.0 **Methods** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **2-5**  
2.1 **Feature Correlation Analysis** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **2**  
2.2 **Principal Component Analysis (PCA)** . . . . . . . . . . . . . . . . . . . . . . . . . . . . **3**  
2.3 **Clustering Analysis with K-Means** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **3**  
2.4 **Docking Score Distribution Analysis** . . . . . . . . . . . . . . . . . . . . . . . . . . . . **4**  
2.5 **Feature Importance in Docking Score Prediction** . . . . . . .. . . . . . . . . . . **5**  
2.6 **Model Performance Evaluation** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **5**

3.0 **Results and Discussion** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **5-7**  
3.1 **Feature Correlation Analysis** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **5**  
3.2 **Principal Component Analysis (PCA) and Chemical Space Exploration** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **6**  
3.3 **Docking Score Distribution Analysis** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **6**  
3.4 **Clustering Analysis of Chemical Space** . . . . . . . . . . . . . . . .. . . . . . . . . . . **6**  
3.5 **Feature Importance in Docking Score Prediction** . . . . . . . . . . . . . . . . . . **7**  
3.6 **Model Performance Evaluation** . . . . . . . . . . . . . .. . . . . . . . . . . . . . . . . . . . **7**

4.0 **Conclusion** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. . . . . . . . . . . . . . . . . . **7-8**

5.0 **Acknowledgments** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **8**

6.0 **References** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **8**

### Introduction

Drug discovery is a complex and resource-intensive process that involves identifying and optimizing potential therapeutic compounds. Computational approaches, particularly molecular docking and machine learning have emerged as powerful tools for accelerating drug discovery by predicting the binding affinity of small molecules to target proteins. Molecular docking predicts how a ligand interacts with a biological target, providing insights into binding affinity, molecular interactions, and pharmacokinetic properties (Shoichet, 2004). However, the effectiveness of docking studies relies heavily on the molecular descriptors used to represent chemical compounds (Lipinski et al., 2001).

In recent years, dimensionality reduction techniques such as Principal Component Analysis (PCA) have been employed to analyze high-dimensional chemical descriptor data, facilitating the identification of key molecular properties that influence docking scores (Jolliffe & Cadima, 2016). PCA enables the visualization of chemical space, where compounds with similar properties cluster together, aiding in the classification and prioritization of potential drug candidates (Todeschini & Consonni, 2009).

Furthermore, unsupervised clustering techniques like K-means clustering allow for the segmentation of compounds into groups based on their physicochemical properties. By clustering compounds into distinct chemical spaces, we can identify trends in molecular behavior and determine which structural features correlate with strong binding affinities (Jain & Nicholls, 2008).

Another crucial aspect of computational drug discovery is machine learning-based prediction models. In this study, we employ a Random Forest regression model to predict docking scores using molecular descriptors. Random Forest is a widely used ensemble learning method known for its robustness against overfitting and its ability to handle high-dimensional data (Breiman, 2001). By evaluating feature importance, we can determine which molecular properties have the greatest impact on docking outcomes, guiding the selection and optimization of promising compounds (Mitchell, 2014).

**Objectives of the Study**

**This research aims to:**

1. Analyze the correlation between chemical descriptors and docking scores using a feature correlation heatmap.
2. Apply PCA to visualize the distribution of compounds in chemical space and identify major variance-contributing features.
3. Perform K-means clustering to classify compounds based on their physicochemical properties.
4. Train a Random Forest model to predict docking scores and identify the most significant molecular descriptors.
5. Evaluate model performance through actual vs predicted docking score analysis.

By integrating PCA, clustering, and machine learning approaches, this study provides valuable insights into the relationship between molecular properties and docking scores, offering a data-driven approach for virtual screening and lead optimization in drug discovery.

# 1. Methods

## All analyses were performed using RStudio 2024.12.1+563 "Kousa Dogwood" Release on Windows 10 (64-bit). The computational workflow was implemented in R, utilizing various libraries for data preprocessing, statistical modeling, and visualization. Data cleaning and manipulation were conducted using dplyr and tidyr, while caret and randomForest were employed for machine learning and predictive modeling. Principal Component Analysis (PCA) and clustering were carried out using factoextra and cluster, and visualizations were generated using ggplot2, ggpubr, gridExtra, and corrplot. The report was compiled using Quarto 1.5.57.

## 2.1 Data Collection and Preprocessing

The dataset was sourced from an open-access repository on GitHub: [Drug Class Structure Dataset](https://raw.githubusercontent.com/HackBio-Internship/2025_project_collection/refs/heads/main/Python/Dataset/drug_class_struct.txt). It contains molecular descriptors for drug-like compounds along with their docking scores against a target protein. The docking scores represent the binding affinity of each molecule, with lower scores indicating stronger interactions (Morris et al., 2009).

To ensure data integrity and optimize analysis, several preprocessing steps were performed. Variable names were standardized using make.names(), and missing values were removed to prevent computational errors. Non-numeric data entries were excluded, and all chemical descriptors were converted into numerical format. Features with zero variance were eliminated to avoid redundancy. Finally, all numerical descriptors were scaled to zero mean and unit variance, ensuring comparability across features.

A correlation analysis was conducted to examine relationships between molecular descriptors. The resulting heatmap (Figure 1) revealed strong interdependencies among features. Notably, molecular weight (MW) and exact molecular weight (MW\_EXACT) exhibited near-perfect correlation, suggesting redundancy. Similarly, TPSA\_NO and TPSA\_NOPS showed significant overlap, indicating that they capture similar molecular characteristics. Additionally, rotatable bond count (RotBondCount) correlated with docking scores, suggesting that molecular flexibility influences binding affinity.

A diagram of a number of numbers

AI-generated content may be incorrect.

**Figure 1.** *Heatmap showing the correlation between molecular descriptors in the dataset. Red areas indicate strong positive correlations, while blue areas represent strong negative correlations. Notable correlations include MW and MW\_EXACT (highly redundant) and TPSA-related features, which are strongly interdependent. The correlation between RotBondCount and docking score suggests a potential relationship between molecular flexibility and binding affinity.*

**2.2 Principal Component Analysis (PCA)**

To explore the underlying structure of the dataset and reduce dimensionality, Principal Component Analysis (PCA) was applied to the chemical descriptors (Jolliffe, 2002). This method allowed for the visualization of chemical space in a lower-dimensional representation while preserving the majority of variance in the data. A scree plot (Figure 3) was used to determine the optimal number of principal components to retain. The first two principal components explained a significant portion of the variance, with PC1 accounting for 40.9% and PC2 for 17.02%. The transformed dataset was subsequently used for clustering analysis.

## 2.3 Clustering Analysis with K-Means

To explore the underlying structure of the dataset and reduce dimensionality, Principal Component Analysis (PCA) was applied to the chemical descriptors (Jolliffe, 2002). This method allowed for visualization of chemical space in a lower-dimensional representation, while preserving most of the dataset’s variance.

A scree plot (Figure 2) was generated to determine the optimal number of principal components to retain. The first two components explained a significant portion of variance, with PC1 accounting for 40.9% and PC2 for 17.02%. This suggested that a 2D projection was sufficient for downstream clustering analysis.

A graph with numbers and a line

AI-generated content may be incorrect.

**Figure 2:** *Scree plot showing the percentage of variance explained by each principal component. PC1 and PC2 capture most of the variance, justifying their use in clustering and further analysis.*

To analyze molecular relationships in PCA space, a dual-panel PCA plot (Figure 3) was generated. The left panel displays compound clustering using K-means (K = 4), while the right panel shows the same PCA projection colored by docking scores. The left panel demonstrates that the dataset naturally groups into four distinct chemical clusters, while the right panel highlights regions associated with strong (low) and weak (high) docking scores.

A screenshot of a graph

AI-generated content may be incorrect.

**Figure 3:** PCA clustering and docking score visualization. The left panel represents K-means clustering of chemical space (K = 4), grouping compounds based on molecular descriptor similarity. The right panel displays the same PCA projection, color-coded by docking scores, where blue indicates strong binding affinity (low scores), and red represents weak interactions (high scores).

**2.4 Docking Score Distribution Analysis**

To examine the overall distribution of docking scores across the dataset, a log-transformed histogram (Figure 4) was generated. The docking scores were log-transformed using log(Docking Score + 1) to address the right-skewed distribution observed in the raw scores. This transformation reduces the impact of extreme values and provides a more normalized representation of docking score frequency.

The histogram reveals a more symmetric distribution of log-transformed docking scores compared to the original, which was heavily skewed. This suggests that, while most compounds exhibit moderate-to-strong binding affinities, the transformation allows for better visualization of score variability, particularly in the presence of high-scoring outliers.

A graph of a log-transformed distribution of docking scores

AI-generated content may be incorrect.

**Figure 4:** *Log-transformed distribution of docking scores. Applying log transformation (Log(Docking Score + 1)) normalizes the previously right-skewed distribution, improving visibility of score variation while reducing the influence of extreme values.*

**2.5 Feature Correlation Analysis**

A correlation analysis was conducted to investigate relationships between molecular descriptors and identify redundant features. The Feature Correlation Heatmap (Figure 1) provides a visual representation of how strongly different descriptors are related to one another.

The heatmap reveals several key patterns:

* Molecular weight (MW) and exact molecular weight (MW\_EXACT) exhibit a near-perfect correlation, indicating that one of these features can be removed to reduce redundancy.
* Topological polar surface area (TPSA\_NO and TPSA\_NOPS) are strongly intercorrelated, suggesting that they capture overlapping molecular properties.
* Rotatable bond count (RotBondCount) shows a moderate correlation with docking scores, implying that molecular flexibility may play a role in ligand binding efficiency.

This correlation analysis was critical for feature selection, ensuring that redundant or highly collinear variables were not included in predictive models, thereby improving model performance and interpretability.

A graph with black dots and white lines

AI-generated content may be incorrect.

**Figure 5:** *Feature Correlation Heatmap of Molecular Descriptors. The heatmap illustrates the correlation between molecular descriptors. Strong positive correlations (red) and negative correlations (blue) indicate interdependencies. MW and MW\_EXACT show high redundancy, while TPSA-related features also exhibit strong correlation. The relationship between RotBondCount and docking scores suggests that molecular flexibility may influence ligand binding efficiency.*

**2.6 Random Forest Regression for Docking Score Prediction**

To predict docking scores based on molecular descriptors, a Random Forest regression model was trained (Breiman, 2001). The model was implemented using the randomForest package, with 100 decision trees to enhance stability and minimize variance.

The model was trained using the preprocessed molecular descriptors as input and docking scores as the output variable. To evaluate feature importance, a ranking of molecular descriptors was extracted from the trained model. The feature importance plot (Figure 6) revealed that TPSA\_NO, FSP3, and MW\_EXACT were the strongest predictors of docking score variations.

A graph with blue bars and white text

AI-generated content may be incorrect.

**Figure 6:** *Feature importance ranking derived from the Random Forest model. TPSA\_NO, FSP3, and MW\_EXACT are the most influential descriptors in docking score prediction.*

**2.7 Model Performance Evaluation**

The predictive performance of the Random Forest model was assessed by comparing actual vs. predicted docking scores (Figure 7). The scatter plot revealed that while the model performed well for low-to-moderate docking scores, it tended to underestimate high-scoring outliers.

This suggests that while molecular descriptors can effectively predict docking scores, additional factors—such as specific protein-ligand interactions—may contribute to extreme docking values.

A graph with a red line and black dots

AI-generated content may be incorrect.

**Figure 7:** *Scatter plot of actual vs. predicted docking scores. The model shows strong performance for lower docking scores but struggles with high-scoring outliers.*

# 2. Results and Discussion

## The results of this study provide insights into the relationships between molecular descriptors, their influence on docking scores, and the predictive capability of machine learning models. By integrating feature correlation analysis, Principal Component Analysis (PCA), K-means clustering, and Random Forest regression, we identified key chemical properties influencing docking performance and evaluated the model’s ability to predict binding affinities.

## Each section below discusses a specific aspect of the analysis, incorporating quantitative findings and visual representations (Figures 1–7) to enhance understanding.

## 2.1 Feature Correlation Analysis

Feature correlation analysis was performed to assess the relationships between molecular descriptors and identify redundant or interdependent features that could impact predictive modeling. The Feature Correlation Heatmap (Figure 1) provides a comprehensive visualization of these interactions.

**Key Observations from the Heatmap**

The heatmap reveals several notable patterns that guided subsequent feature selection and model optimization:

* Molecular weight (MW) and exact molecular weight (MW\_EXACT) exhibit a near-perfect correlation (r ≈ 1.0), confirming redundancy. Since MW\_EXACT is a precise measure of molecular weight, MW can be excluded from predictive modeling without losing relevant information.
* Topological polar surface area (TPSA\_NO) and TPSA\_NOPS are strongly intercorrelated, suggesting that they capture overlapping molecular properties related to hydrogen bonding and polarity. Retaining only one of these descriptors may suffice for docking score prediction.
* Rotatable bond count (RotBondCount) shows a moderate correlation with docking scores, implying that molecular flexibility plays a role in ligand binding efficiency.

These findings provided a data-driven approach to feature selection, ensuring that only the most informative molecular descriptors were used in dimensionality reduction and predictive modeling.

**3.2 Principal Component Analysis (PCA) and Chemical Space Exploration**

To investigate the structural diversity of compounds and facilitate clustering analysis, Principal Component Analysis (PCA) was applied to the filtered set of molecular descriptors. PCA is a widely used technique for dimensionality reduction, allowing complex datasets to be visualized in lower-dimensional space while preserving the majority of variance (Jolliffe, 2002).

Explained Variance and Optimal Component Selection

The scree plot (Figure 2) displays the proportion of variance explained by each principal component. The first two components accounted for 57.9% of the total variance (PC1 = 40.9%, PC2 = 17.02%), indicating that a 2D projection captures most of the structural variation in the dataset.

**PCA-Based Visualization of Chemical Space**

A dual-panel PCA scatter plot (Figure 3) was generated to provide a detailed view of compound distribution:

* The left panel presents K-means clustering (K = 4), which identified four distinct chemical clusters, suggesting that compounds share similar molecular features within each group.
* The right panel maps docking scores onto PCA space, highlighting that high-affinity binders (low docking scores) tend to cluster in specific regions, reinforcing the role of structural properties in ligand-protein interactions.

These findings validate PCA as an effective method for reducing complexity while retaining chemically meaningful relationships among compounds.

**3.3 Docking Score Distribution Analysis**

To gain insights into the overall distribution of docking scores, a log-transformed histogram (Figure 4) was generated. Raw docking scores were highly skewed, making it difficult to discern trends within the dataset. A log transformation (Log(Docking Score + 1)) was applied to normalize the distribution and improve visibility.

**Key Insights from the Log-Transformed Distribution**

* The transformed histogram reveals a more symmetric distribution, improving visualization of score variability.
* The majority of compounds exhibit moderate-to-strong binding affinities, suggesting that a large portion of the dataset contains viable drug candidates.
* The presence of high-scoring outliers suggests potential steric clashes, unfavorable electrostatic interactions, or poor ligand complementarity in docking simulations.

These results highlight the need for further molecular optimization to enhance binding affinity for lead compounds.

**3.4 Clustering Analysis of Chemical Space**

## To further investigate structure-activity relationships, K-means clustering (K = 4) was applied to the PCA-transformed dataset. This approach enabled the identification of structurally similar compounds, which were then analyzed for differences in docking scores.

## Cluster Distribution and Docking Scores (Figures 3 & 5)

## Figure 3 (left panel) confirms the presence of four distinct chemical clusters, suggesting that molecular similarity drives natural grouping.

## Boxplots (Figure 5) reveal that Cluster 2 exhibits the lowest median docking score, indicating that compounds within this cluster share structural features favorable for binding.

## This clustering approach prioritizes chemical scaffolds that demonstrate strong protein-ligand interactions, aiding in drug discovery efforts.

**3.5 Feature Importance in Docking Score Prediction**

To predict docking scores based on molecular descriptors, a Random Forest regression model was trained. Feature importance analysis was conducted to identify which molecular properties contributed most significantly to docking performance.

Top Predictors of Docking Score (Figure 6)

The feature importance plot (Figure 6) ranks descriptors based on their influence on docking scores:

1. TPSA\_NO – A measure of hydrogen bonding capacity, indicating the role of polar interactions in docking efficiency.
2. FSP3 – Represents the fraction of sp3-hybridized carbons, linking molecular flexibility to enhanced binding.
3. MW\_EXACT – Highlights the influence of steric effects and molecular stability on docking interactions.

These results confirm that polar interactions, molecular flexibility, and steric considerations are key determinants of docking success.

**3.6 Model Performance Evaluation**

The predictive performance of the Random Forest model was assessed by comparing actual vs. predicted docking scores (Figure 7).

Evaluation of Predictive Accuracy

* The model accurately predicts low-to-moderate docking scores, suggesting that molecular descriptors effectively capture binding trends.
* Underestimation of high-scoring outliers suggests that additional molecular properties—such as conformational flexibility and solubility—may need to be incorporated to refine predictions.
* The overall agreement between predicted and actual values supports the use of machine learning in docking score estimation.

Future improvements may involve hyperparameter tuning, ensemble learning, or deep neural networks to enhance predictive performance.

**4.0 Conclusion**

This study demonstrates how dimensionality reduction, clustering, and machine learning techniques can be leveraged to analyze chemical space and predict docking scores. By integrating Principal Component Analysis (PCA), K-means clustering, feature correlation analysis, and Random Forest regression, we identified key molecular descriptors influencing docking affinity and evaluated the predictive power of machine learning models. The findings provide valuable insights into structure-activity relationships (SAR) and offer a data-driven framework for virtual screening and lead optimization in drug discovery.

Key Findings and Contributions

1. Feature correlation analysis revealed significant interdependencies among molecular descriptors. Molecular weight (MW) and MW\_EXACT were highly redundant, while TPSA\_NO and TPSA\_NOPS displayed strong correlation, highlighting the need for careful feature selection to avoid overfitting and redundancy in predictive modeling.
2. PCA demonstrated that a reduced set of molecular descriptors could explain most of the variance in chemical space. The first two principal components accounted for 57.9% of the total variance, justifying their use for clustering analysis and visualization.
3. K-means clustering (K = 4) revealed distinct molecular groupings, with one cluster (Cluster 2) containing compounds that exhibited stronger binding affinities. This finding suggests that certain molecular properties—such as hydrogen bonding potential and molecular flexibility—are associated with improved docking performance.
4. Docking score distribution analysis indicated that the majority of compounds exhibited moderate-to-strong binding affinities, while a small subset of high-scoring outliers may represent unfavorable binding interactions due to steric clashes or poor complementarity.
5. Machine learning-based docking score prediction using Random Forest regression identified the most influential molecular descriptors. TPSA\_NO, FSP3, and MW\_EXACT were the top predictors, confirming the role of hydrogen bonding capacity, molecular flexibility, and steric factors in ligand-protein interactions.
6. Model evaluation revealed strong predictive capability for low-to-moderate docking scores, but the model underestimated high-scoring outliers, suggesting the need for further refinement, such as incorporating conformational dynamics, ligand solubility, and entropic effects.

Implications for Drug Discovery and Molecular Design

The findings of this study have several implications for rational drug design and computational screening approaches:

* Virtual Screening Optimization – By identifying molecular descriptors that strongly influence docking scores, computational filtering of drug candidates can be improved, reducing the number of false positives in virtual screening pipelines.
* Molecular Scaffolding and Lead Optimization – Clustering analysis suggests that certain molecular scaffolds are associated with strong protein-ligand interactions. Future studies can focus on derivative compounds within high-affinity clusters to enhance potency.
* Feature Engineering for AI-based Drug Design – This study highlights the importance of careful feature selection in machine learning models for drug discovery. The insights gained from feature correlation analysis can guide the development of improved descriptors that capture more complex interaction patterns.

Limitations and Future Directions

While the study provides valuable insights into molecular features influencing docking scores, there are several areas that require further investigation:

* Incorporating Additional Molecular Features – The model could be enhanced by integrating 3D molecular descriptors, solubility predictions, or ligand conformational flexibility to capture nonlinear binding interactions.
* Deep Learning Approaches – Advanced models such as Graph Neural Networks (GNNs) or Transformer-based architectures may provide superior predictive accuracy by capturing spatial and quantum-level interactions.
* Experimental Validation – While docking scores serve as an important computational metric, wet-lab validation (e.g., binding assays, crystallography studies) is essential to confirm the predicted ligand-protein interactions.

Conclusion and Outlook

This study highlights the power of data-driven approaches in drug discovery, demonstrating how machine learning and statistical techniques can reveal meaningful trends in ligand binding affinity. By integrating PCA, clustering, and Random Forest regression, we identified key molecular descriptors that influence docking scores, paving the way for more efficient virtual screening and lead optimization.

As computational tools continue to evolve, future research should focus on combining physics-based modeling with deep learning techniques to further enhance binding affinity predictions. Additionally, expanding these analyses to larger compound libraries and multiple protein targets will improve the generalizability of predictive models, bringing us closer to an era of AI-driven molecular design and precision drug discovery.

### Acknowledgments

The authors would like to express their gratitude to HackBio for providing this internship and training program, which has played a crucial role in advancing skills in bioinformatics, data analysis, and computational modeling. The opportunity to work with real-world datasets and apply machine learning techniques to drug discovery has been invaluable in enhancing expertise in computational biology and cheminformatics.

Additionally, appreciation is extended to the HackBio mentors, peers, and collaborators for their continuous support, insightful discussions, and guidance throughout the project. This experience has contributed significantly to the development of bioinformatics competencies and has reinforced the importance of data-driven approaches in drug discovery.

### References

1. Bender, A., & Cortes-Ciriano, I. (2021). Artificial intelligence in drug discovery: What is real and what is hype? *Genome Biology, 22*(1), 1-14. https://doi.org/10.1186/s13059-021-02324-2
2. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324
3. Gaudelet, T., Day, B., Jansen, J., Liu, G., & Zang, X. (2020). Utilizing deep learning and generative models for drug design. *Nature Reviews Drug Discovery, 19*(8), 527–542. https://doi.org/10.1038/s41573-020-0073-z
4. Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 28*(1), 100–108. https://doi.org/10.2307/2346830
5. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011
6. Jolliffe, I. T. (2002). Principal component analysis (2nd ed.). Springer. https://doi.org/10.1007/b98835
7. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2
8. Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer. https://doi.org/10.1007/978-1-4614-6849-3
9. Lever, J., Krzywinski, M., & Altman, N. (2017). Principal component analysis. *Nature Methods, 14*(7), 641–642. https://doi.org/10.1038/nmeth.4346
10. Lipinski, C. A. (2004). Lead- and drug-like compounds: The rule-of-five revolution. *Drug Discovery Today: Technologies, 1*(4), 337–341. https://doi.org/10.1016/j.ddtec.2004.11.007
11. Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., & Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry, 30*(16), 2785–2791. https://doi.org/10.1002/jcc.21256
12. Todeschini, R., & Consonni, V. (2009). Molecular descriptors for chemoinformatics. Wiley-VCH. https://doi.org/10.1002/9783527628766

1. [↑](#footnote-ref-1)