Balancing Interpretability and Predictive Power: Machine Learning Approaches for Mutagenicity Prediction

Miko F. Mallari
The University of California, San Francisco, San Francisco, CA, USA
Schools of Pharmacy and Medicine
Department of Bioengineering and Therapeutic Sciences
Current Address: Genentech Hall, 600 16th St, San Francisco, CA 94158
Correspondence should be addressed to M.F.M (miko.mallari@ucsf.edu)

Accurate prediction of chemical mutagenicity is essential for prioritizing compounds in early drug discovery and reducing reliance on resource-intensive laboratory testing. In this study, multiple machine learning algorithms—Random Forest, k-Nearest Neighbors, Support Vector Machine, Logistic Regression, and XGBoost—were trained and evaluated on the AMES Mutagenicity Dataset to predict mutagenic potential from molecular fingerprints. Two primary feature sets were investigated: a small set of RDKit-derived descriptors (Morgan fingerprints, molecular weight, and heavy atom count) and a more extensive set of CDK-based fingerprints generated from the SMILE strings via PaDELpy. While the simpler RDKit features were interpretable, they did not produce strong predictive performance (maximum ROC AUC ~65.5%) and were prone to overfitting. In contrast, the CDK fingerprints improved generalizability, with a kNN classifier showing minimal overfitting and improved ROC AUC scores, though at the cost of reduced interpretability. These findings demonstrate a trade-off between interpretability and predictive accuracy, underscoring the need for careful feature selection and suggesting that incorporating more established, interpretable fingerprint types, external validation, and computational scalability assessments will be crucial for future advances in mutagenicity prediction models.

Introduction

Accurately predicting mutagenicity in chemical substances is crucial for assessing potential health risks and informing regulatory decisions. [1] While valuable, traditional laboratory-based methods, such as the Ames test, can be time-consuming and resource-intensive. To address these limitations, computational approaches, known as in silico screening, have emerged as powerful tools for rapid and efficient mutagenicity prediction.

This study employs a combination of machine learning algorithms from SciKit-Learn's (SKL) repository to develop predictive models: [2]

Random Forest Classifier (RFC): A Random Forest is a learning method that constructs multiple decision trees on different subsets of the data and then aggregates their predictions to improve accuracy and reduce overfitting. According to SKL documentation, RFC uses averaging to enhance predictive performance and robustness. In this study, it is employed to classify compounds as mutagenic or non-mutagenic by leveraging the combined insights from numerous decision trees to identify structural patterns linked to mutagenicity. [2]

k-Nearest Neighbors (kNN): The kNN algorithm is a non-parametric supervised learning method. As described by SKL, kNN assigns a data point to the most common class among its nearest neighbors. In this project, kNN evaluates the similarity between molecular fingerprints to classify compounds, allowing the identification of mutagenic patterns through nearest-neighbor comparisons. [2]

Support Vector Machine (SVM): SVM aims to find the optimal hyperplane that separates data points into different classes. By maximizing the margin between the classes, SVM can achieve high accuracy, especially in high-dimensional spaces. [2]

Logistic Regression: Logistic regression class is a linear model used for classification tasks that predicts the probability of a sample belonging to a given class. It utilizes a logistic function to map linear combinations of input features to a probability value between 0 and 1. [2]

XGBoost (Extreme Gradient Boosting): This method builds multiple decision trees sequentially, with each tree focusing on correcting the errors of the previous ones. In the context of this study, XGBoost incrementally improves upon previous decision trees to enhance predictive power.[3]

These models are trained on datasets of chemical compounds with known mutagenicity, leveraging molecular fingerprints derived from chemical structures. Specifically, we utilize RDKit Morgan Fingerprints

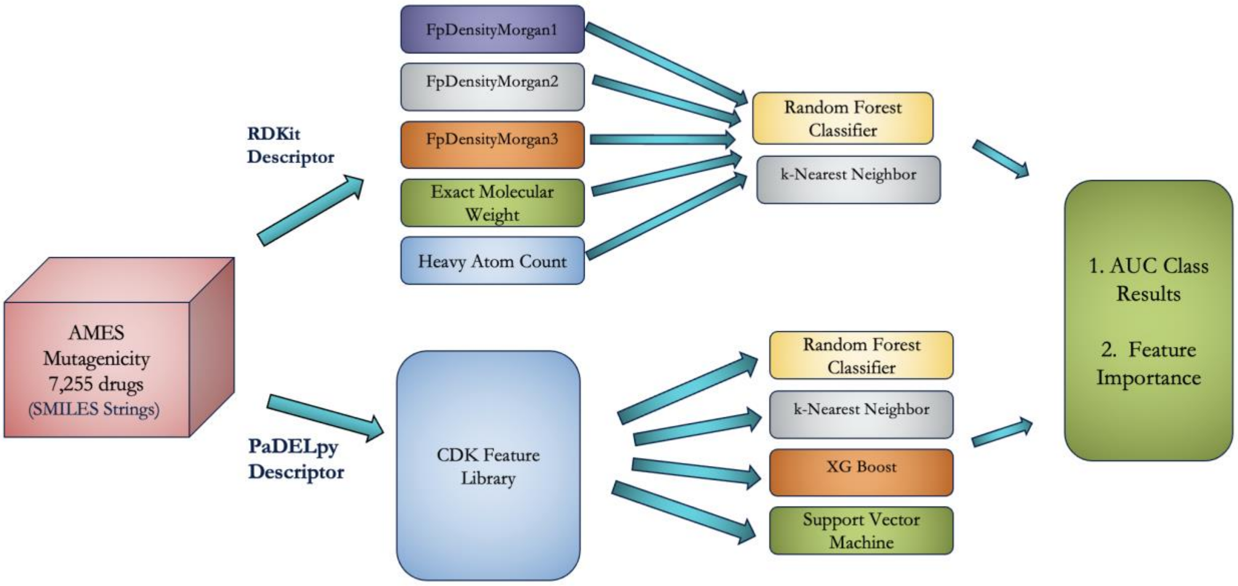


Figure 1: Graphical overview of workflow. The AMES mutagenicity Dataset was divided into training, validation, and test sets to develop a binary classification model for predicting mutagenicity. Feature extraction involved generating three morgan footprints, molecular weight, and heavy atom count via RDKit to capture basic structural and physicochemical properties. Subsequent experiments employed PaDELpy to convert SMILES into CDK fingerprints. Multiple machine learning algorithms, including kNN, Logistic Regression, SVM, XGBoost, and RFCs were trained and evaluated using ROC AUC.

generated using the open-source cheminformatics toolkit RDKit and PaDEL-Descriptor, a Java-based software for calculating molecular descriptors. [2,4]

Molecular fingerprints, such as RDKit Morgan Fingerprints, encode information about the molecular structure, including atom types, bond types, and spatial arrangements. By comparing these fingerprints, the similarity between molecules can be assessed. This similarity measure is crucial for machine learning algorithms to learn patterns and make accurate predictions. [4]

By combining these powerful machine learning algorithms with informative molecular fingerprints, we aim to develop robust and accurate models for predicting mutagenicity. These models can be valuable tools for prioritizing chemicals for further testing, reducing the need for extensive laboratory experiments, and ultimately contributing to safer chemical design and use.

Methods

Data Source and Splits

This study utilized the AMES Mutagenicity Dataset sourced from Therapeutics Data Commons (pyTDC), comprising 7,255 compounds labeled as mutagenic or non-mutagenic (55% and 45%, respectively). [5] The dataset was randomly divided into training (70%), validation (20%), and test (10%) sets. The training set was used to fit the models, the validation set guided hyperparameter selection, and the held-out test set evaluated final model performance.

Feature Extraction

For the initial modeling approach, molecular descriptors were extracted from chemical structures represented as simplified molecular line entry system (SMILES) representations of the compounds. Five features were selected: three Morgan fingerprints (fp1, fp2, fp3), molecular weight, and heavy atom molecular weight. Morgan fingerprints were generated using RDKit, capturing the structural neighborhood around each atom. No additional scaling or feature transformations were performed for these initial features.

Model Selection and Algorithms

Two primary approaches were tested in the initial phase: a RFC and a kNN classifier. Both the RFC and the kNN models were implemented using RDKit to leverage direct integration with molecular descriptors and provide feature importance measures. The choice of these algorithms was guided by prior references to a reference paper, where kNN exhibited strong performance, and RFC provided a robust baseline and interpretability through feature importance scores. [6]

Additional Analysis

After the initial evaluation, additional feature extraction was performed using PaDELpy to generate Chemistry Development Kit (CDK) fingerprints. These fingerprints produce a binary vector where each position corresponds to a predefined structural feature. The presence of a feature is encoded as 1; absence as 0. This representation captures a broader range of substructure patterns, enabling a more comprehensive structural characterization.

With the expanded feature set (CDK fingerprints), multiple machine learning models were assessed, including: kNN, Logistic Regression, SVM, XGBoost, and RFC. Each model's performance on the validation set guided the selection and tuning of hyperparameters.

Hyperparameter Tuning

Hyperparameter optimization was performed for each algorithm in the models examined based on editable parameters as presented in the documentation of each algorithm. Further optimization included an assessment of varied learning rates, each tested at learning rates of [0.001, 0.01, 0.1, 1, 10, 100]. Each configuration's performance was evaluated using Area Under the Receiver Operating Characteristic curve (ROC AUC) at Test, Training, and Validation datasets.

Performance Metric and Feature Importance

The primary evaluation metric was the ROC AUC. This threshold-independent metric effectively captures the trade-off between sensitivity and specificity, making it suitable for binary classification tasks with imbalanced classes. The ROC AUC was computed for the training, validation, and test sets to monitor overfitting and to identify the final, generalizable model configuration. For both analyses, feature importance was calculated to provide a relative measure of each feature's contribution to reducing impurity across the models.

Results and Discussion

The seven models, Test, Train, and Validation ROC AUC scores, were assessed to cross-evaluate the models against themselves and each other. Overfitting was evaluated by calculating the difference between validation scores and test scores.

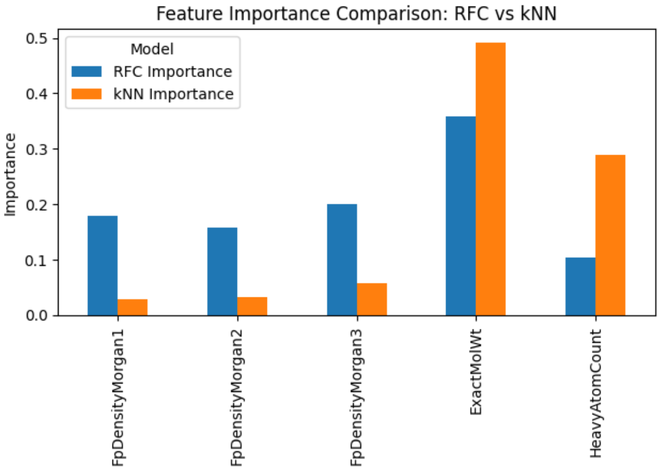
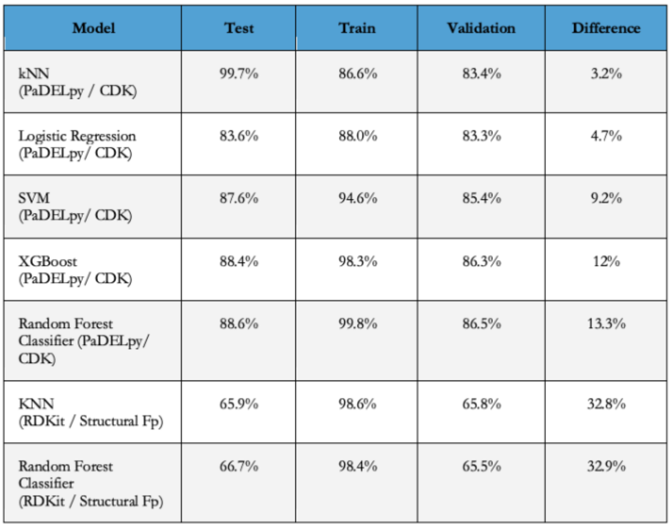


Figure 2: RDKit fingerprints and RFC and kNN feature importance analysis. Feature importance of the physical and structural features for classifying mutagenic SMILES. The horizontal bar chart shows the relative importance of each feature as determined by RFC and kNN. Higher importance values indicate a stronger contribution of the feature to the model's classification performance.



Model	Test	Train	Validation	Difference
kNN (PaDELpy / CDK)	99.7%	86.6%	83.4%	3.2%
Logistic Regression (PaDELpy/ CDK)	83.6%	88.0%	83.3%	4.7%
SVM (PaDELpy/ CDK)	87.6%	94.6%	85.4%	9.2%
XGBoost (PaDELpy/ CDK)	88.4%	98.3%	86.3%	12%
Random Forest Classifier (PaDELpy/ CDK)	88.6%	99.8%	86.5%	13.3%
KNN (RDKit / Structural Fp)	65.9%	98.6%	65.8%	32.8%
Random Forest Classifier (RDKit / Structural Fp)	66.7%	98.4%	65.5%	32.9%

Table 1: Summary of model performance based on ROC AUC scores. Test, Train, and Validation scores based on ROC AUC test. Difference was calculated by subtracting Test from Validation scores.

The results of this study highlight the critical role of feature selection and representation when predicting mutagenicity from molecular structures. With the initial feature of three Morgan fingerprints, molecular weight, and heavy atom count, feature importance analyses provided a degree of interpretability. Each feature's contribution to reducing entropy from RDKit fingerprints in RFC and kNN was observed, and molecular weight resulted in the most important feature of the 5 (Figure 2). This theoretically allowed insight into which structural characteristics might correlate with mutagenic outcomes. However, despite the ability to identify structural patterns, the selected features did not yield strong indicators of mutagenicity, with an overall AUC ROC score of 65.5%. In other words, while these features captured certain aspects of molecular structure and physical properties, they did not consistently translate into robust predictive signals.

PaDELpy derived CDK-based fingerprints were used to continue this investigation. CDK is less interpretable at the individual feature level since specific substructures are not directly mapped to a known key. However, it provided many more features to (>1000) to train models. While losing some interpretability, these fingerprints encapsulate various chemical patterns, potentially enhancing the model's ability to detect subtle structural signals. Nevertheless, the trade-off is clear: the standard CDK fingerprint's ability to provide additional information to the various models resulted in a clear advantage to the five structural features in the original model for developing predicting mutagenicity.

The CDK library of features provided a richer fingerprint context to fit the models. This approach, however, does not provide additional information for the mechanistic understanding of why certain molecules are flagged as mutagenic. While raw predictive performance may improve with more complex or comprehensive fingerprints, interpretability remains challenging when the features cannot be easily mapped back to specific chemical substructures.

Overfitting emerged as a key concern in evaluating model stability and generalizability. Models relying on structural fingerprint features alone exhibited substantial discrepancies between training and validation ROC AUC scores, at times exceeding a 30% difference. While the RFC model using PaDELpy and CDK fingerprints achieved the highest score, it also displayed the greatest degree of overfitting among the PaDELpy-based models, with a 13.3% gap between training and validation performance. In contrast, the kNN model employing PaDELpy CDK fingerprints showed the smallest train-validation difference (3.2%), indicating better generalization. Similarly, the Logistic Regression model using the same fingerprint set demonstrated low variance across all data splits, highlighting the stability of this approach. These findings suggest that although CDK-based features may be less interpretable, they can yield models with superior generalizability. Notably, the CDK combinations align with existing literature, where kNN and SVM models performed well when paired with CDK-based MACCS and PubChem fingerprints. [6]

This alignment with previous studies showed that certain fingerprint types—MACCS, PubChem, and other widely recognized fingerprints—served as robust and interpretable descriptors for mutagenicity-related endpoints. The reference paper's emphasis on kNN and SVM models using MACCS and PubChem supports the conclusion that selecting well-established, interpretable feature sets can be as important as selecting a classifier. [6] Although the present study's feature sets did not yield a strongly predictive or easily interpretable framework for mutagenicity in a single model, these findings suggest a clear path forward.

Future Directions

Future work will emphasize striking a balance between interpretability and predictive power. Exploring alternative fingerprints, such as MACCS keys and PubChem fingerprints, offers the potential to enhance predictive performance while improving mechanistic understanding. Incorporating external validation datasets, as employed in reference studies, will further verify the generalizability of the models beyond the AMES Mutagenicity dataset. Complexity analyses, including Bachmann-Landau (Big-O) notation assessments, could provide valuable insights into the computational scalability and efficiency of these approaches for more extensive chemical libraries. [7] While this study has laid the groundwork by incorporating additional fingerprints for exploration, further analysis is required to refine the trade-offs among interpretability, performance, and generalizability in mutagenicity prediction models.

References

1. Ames, B. N., Durston, W. E., Yamasaki, E., & Lee, F. D. (1973). Carcinogens are Mutagens: A Simple Test System Combining Liver Homogenates for Activation and Bacteria for Detection. *Proceedings of the National Academy of Sciences*, *70*(8), 2281–2285. https://doi.org/10.1073/pnas.70.8.2281
2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (n.d.). Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*.
3. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. ACM.
4. RDKit: Open-source cheminformatics; http://www.rdkit.org
5. Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., & Xiao, C. (2021). *Therapeutics Data Commons: Machine learning datasets and tasks for drug discovery and development*. *arXiv*. https://arxiv.org/abs/2102.09548
6. Xu, C., Cheng, F., Chen, L., Du, Z., Li, W., Liu, G., Lee, P. W., & Tang, Y. (2012). In silico Prediction of Chemical Ames Mutagenicity. *Journal of Chemical Information and Modeling*, *52* (11), 2840–2847. https://doi.org/10.1021/ci300400a
7. Corman, T.H, Leiserson, C.E., Rivest, R.L., & Stein, C. (2022). *Introduction to Algorithms*. (4th ed.). MIT Press.

Acknowledgments

This project was made possible by a group of first-year graduate students and their professors at the University of California, San Francisco's Artificial Intelligence and Computational Drug Discovery and Development program. Special thanks to Issac Yeoh for workflow and technical guidance, Jason Rivera for note-taking and time management, Xiao Guo for project and group management, our teaching assistants Jeff Lin and Philip Yuan for their patience and commitment to support this cohort, and Drs. Miao and Sun, thank you for your mentorship and for sharing your knowledge and passion for machine learning.