RESEARCH ARTICLE

Predicting Chemical Biodegradability for Sustainable Chemical Manufacturing: A Machine Learning Approach Using 3D Molecular Descriptors

Alaa M Elsayad[1*], Hassan Yousif Ahmed[1], Khaled A Elsayad[2], Ammar Elyas Babiker Hassan[3], Mustafa Mohammed Hassan Mustafa[4], Akhtar Nawaz Khan[5], Arif Abdelwhab Ali[6], Sahar A. Mokhtar[7]

[1] Department of Electrical Engineering, College of Engineering in Wadi Alddawasir, Prince Sattam Bin Abdulaziz University, Wadi Alddawasir 11991, Saudi Arabia

[2] Pharmacy Department, Cairo University Hospitals, Cairo University, Cairo 11662, Egypt

[3] Electronic Systems Software Engineering Department, Faculty of Engineering, University of Science and Technology, 12211, Sudan

[4] Computer Engineering Department, College of Engineering, Future University, 12211, Sudan

[5] Department of Electrical Engineering, University of Engineering & Technology, Peshawar, Jalozai Campus, 25120, Pakistan

[6] Computer Science Department, Faculty of Mathematical & Computer Science, University of Gezira, 21111, Sudan

[7] Computer and Systems Department, Electronic Research Institute, Cairo 11662, Egypt

Abstract: Achieving sustainable cities and promoting responsible consumption require innovative approaches to chemical design and manufacturing. Precise prediction of chemical biodegradability is crucial for evaluating environmental concerns and facilitating the transition towards green chemistry. This study investigates the effectiveness of ten distinct groups of three-dimensional (3D) molecular descriptors for classifying compounds with rapid biodegradability. The Merck molecular force field (MMFF94s) was used to compute descriptors and generate 3D conformations for a dataset of chemical compounds. The dataset underwent rigorous preprocessing, including feature selection, outlier management, and scaling. Support Vector Machines (SVMs) were tested alongside three tree-based ensemble learning algorithms: Extreme Gradient Boosting (XGBoost), Gradient Boosting Machine (GBM), and Random Forest. Bayesian optimization was employed to optimize model hyperparameters and enhance cross-validated Area Under the Receiver Operating Characteristic Curve (AUC). The GETAWAY descriptors, 3D autocorrelation descriptors, and 3D-MoRSE descriptors consistently demonstrated superior performance compared to other descriptors across all machine learning models. An SVM model trained on 3D autocorrelation descriptors achieved the highest prediction accuracy (0.88), sensitivity (0.83), specificity (0.91), F1-score (0.82), Cohen's Kappa statistic (0.74), and an AUC of 0.93 on an independent test set. Advanced analytical techniques, including Permutation Feature Importance (PFI), SHapley Additive exPlanations (SHAP), and partial dependency plots (PDP) were utilized to identify the most influential 3D autocorrelation descriptors. The findings of this study demonstrate that 3D molecular descriptors, particularly 3D autocorrelations, play a critical role in developing accurate and interpretable models for predicting chemical biodegradability. These models contribute significantly to the advancement of green chemical design and the development of effective regulatory policies that support the objectives of SDG 11 (Sustainable Cities and Communities) and SDG 12 (Responsible Consumption and Production). By fostering sustainable chemical manufacturing practices, we can create healthier and more resilient urban environments while minimizing the environmental impact of human activities.

Keywords: Biodegradability, 3D molecular descriptors, SVM, XGboost, gradient boosting, random forest permutation feature importance, SHAP, QSAR, environmental risk assessment, sustainable chemistry

Correspondence to: Alaa M Elsayad, Department of Electrical Engineering, College of Engineering in Wadi Alddawasir, Prince Sattam Bin Abdulaziz University, Wadi Alddawasir 11991, Saudi Arabia; E-mail: a.elsayyad@psau.edu.sa

Received: December 20, 2024; **Accepted:** December 24, 2024; **Published Online:** December 30, 2024

Citation: Alaa, M. E., Hassan, Y. A., Khaled, A. E., Ammar, E. B. H., Mustafa, M. H. M., Akhtar, N. K., Arif, A. A. Sahar, A. M., 2024. Predicting Chemical Biodegradability for Sustainable Chemical Manufacturing: A Machine Learning Approach Using 3D Molecular Descriptors. *Applied Environmental Biotechnology*, 9(2): 76-86. http://doi.org/10.26789/AEB.2024.02.009

1 Introduction

The growing global demand for sustainable chemical solutions is driven by mounting concerns over environmental pollution and the necessary need for safer alternatives. Accurate identification of readily biodegradable (RB) chemicals is crucial for environmental risk assessment, minimizing pollution, and promoting the development of eco-friendly products. However, traditional experimental methods for assessing biodegradability often prove time-consuming, expensive, and limited by their laboratory-based nature, which may not fully reflect real-world environmental conditions (Abdullah and Abdulazeez, 2021). Computational Quantita-

tive Structure-Activity Relationship (QSAR) models, particularly those leveraging molecular descriptors and machine learning algorithms, offer a promising alternative for faster and more efficient biodegradability prediction (Ahmadi et al., 2023) and (Bahia et al., 2023). Molecular descriptors encode information about a molecule's structure and properties, enabling the development of predictive models that can classify chemicals as readily biodegradable or non-readily biodegradable. While 2D descriptors have been widely used in biodegradability prediction, recent advancements in computational chemistry and machine learning have made 3D descriptors increasingly accessible and valuable (Consonni and Todeschini, 2010). Previous studies have demonstrated the utility of 3D descriptors for modeling various toxicological and environmental endpoints. However, the comparative performance of different 3D descriptor groups for the classification of readily biodegradable compounds remains understudied (Consonni and Todeschini, 2010).

In this study, we investigate the performance of ten distinct 3D molecular descriptor groups for classifying readily biodegradable compounds (Haddouchi and Berrado, 2019, Lombardo et al., 2014, Chinedu et al., 2013). These descriptors can capture detailed information about the spatial arrangement and interactions of atoms within a molecule. We generated 3D molecular conformations for a large dataset of compounds and calculated the 10 different 3D descriptor groups, including: 3D Matrix-Based, 3D Autocorrelations, 3D Atom Pairs, Chemically advanced template search (CATS), Geometrical, Radial distribution function (RDF), 3D Molecular Representation of Structures based on Electron diffraction (3D-MoRSE), Weighted Holistic Invariant Molecular (WHIM), GEometry, Topology, and Atom-Weights AssemblY (GETAWAY), Weighted Holistic Atom Localization and Entity Shape (WHALES). The predictive powers of different 3D descriptor groups were explored through a rigorous process involving feature selection, anomaly treatment, cross-validation, model optimization, and evaluation using various machine learning algorithms. Our goal is to identify the most effective 3D descriptor groups for accurately predicting readily biodegradability, paving the way for more efficient and reliable methods for identifying environmentally friendly chemicals. We compared the performance of three tree-based ensemble models (XGBoost, Gradient Boosting, and Random Forest) and Support Vector Machines (SVM) for the biodegradability classification task, optimizing the hyperparameters of each model using Bayesian optimization. Our findings demonstrate the utility of GETAWAY, 3D autocorrelation, and 3D-MoRSE performed best across all machine learning models outperforming other groups. The insights gained from this study can inform the development of accurate and interpretable computational models for the environmental risk assessment of chemicals, ultimately supporting more sustainable chemical design and regulation.

2 Data Acquisition and Descriptor Extraction

The dataset used in this study was sourced from literature and included the necessary information for the analysis (Lu et al., 2023). Specifically, the dataset contained the CAS Registry Numbers (CAS-RN), SMILES codes, and biodegradation classifications (Readily Biodegradable or Non-Readily Biodegradable) for a set of chemical substances. The CAS-RN is a unique identifier that distinguishes individual chemical compounds, even when multiple names exist for the same substance. The SMILES code, on the other hand, is a line notation that represents the chemical structure of a molecule. To ensure the quality of the dataset, the SMILES codes were checked and canonicalized using the alvaMolecule software (Mansouri et al., 2013), and any duplicate entries were removed. This process resulted in a final dataset of 1717 unique chemical records, consisting of 545 Readily Biodegradable (RB) and 1172 Non-Readily Biodegradable (NRB) compounds. The entire data processing and modeling workflow was executed within the KNIME platform, utilizing a multi-step approach. Starting with SMILES codes, we generated 3D molecular structures using the RDKIT node and the MMFF94s force field. To account for molecular flexibility, up to 10 conformers were generated per molecule. Recognizing the sensitivity of 3D descriptors to conformational variation—ranging from highly sensitive descriptors like 3D Matrix-Based, 3D Autocorrelations, 3D Atom Pairs, and CATS 3D to less sensitive descriptors such as Geometrical, RDF, 3D-MoRSE, WHIM, GETAWAY, and WHALES—we employed a strategy of generating multiple conformations and averaging descriptor values. This approach ensured a more robust and representative feature vector for each compound. To capture a comprehensive set of 3D structural features relevant to biodegradability, we calculated ten distinct 3D molecular descriptor groups using the alvaDesc calculator node within KNIME. Figure 1 shows descriptor generation and feature selection steps.

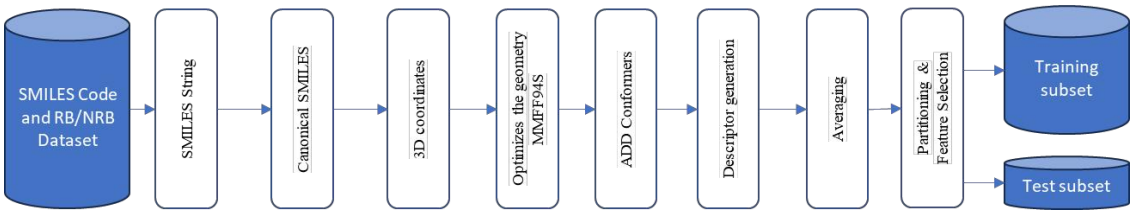


Figure 1. Descriptor generation and feature selection steps.

Table 1 provides a summary of each descriptor group, including a brief description and the number of descriptors within each group. Recognizing the inherent flexibility of molecules and the sensitivity of 3D descriptors to conformational variations, we implemented a multi-conformer averaging strategy. This involved generating multiple conformations for each molecule and then averaging the descriptor values across these conformers. This approach mitigates the potential bias introduced by a single, potentially arbitrary conformation, leading to a more reliable and generalizable representation of the molecule's 3D structure. By accounting for conformational flexibility, we can develop more robust and accurate predictive models for classifying readily biodegradable chemicals.

Figure 2 provides a statistical summary of the Information Gain (IG) values of each 3D descriptor group. The table includes the number of descriptors, maximum IG, mean IG, standard deviation (STD) of IG values, and the overall sum of the IG values of each group (Mauri and Bertola, 2022). Notably, the 3D autocorrelation descriptor group stands out as the most informative for predicting the biodegradability of the compounds. This group contains the feature with the maximum IG of 0.206, the highest mean IG of 0.085, and the largest standard deviation of 0.052. These statistics indicate that the 3D autocorrelation descriptors capture the most relevant structural information for distinguishing readily biodegradable from non-readily biodegradable chemicals. Figure 2 further illustrates the maximum IG achieved by the different 3D descriptor groups. This visual representation reinforces the finding that the 3D autocorrelation descriptors possess the most predictive power among the evaluated groups.

3 Preprocessing and Feature Selection

As a preprocessing step, any constant or near-constant descriptors (with variance less than 0.01) were removed from the dataset to avoid redundant or uninformative features. The remaining dataset was then normalized with z-score normalization (Gaussian) and partitioned into training and test subsets, with an 80:20 split, respectively. To identify the most relevant and uncorrelated features for the subsequent modeling and analysis, a multi-step feature selection process was implemented. First, for each 3D descriptor group, the descriptors were ranked according to their Information Gain (IG) values. Descriptors with an IG less than 0.01 of the group's maximum IG were excluded from further consideration. This step ensured that the models would focus on the most predictive structural characteristics of the molecules. Next, the remaining descriptors were subjected to correlation analysis. Descriptors with a correlation exceeding 95% with other selected descriptors were removed, retaining only one representative from each highly correlated group. This helped to minimize redundancy and multicollinearity among the input features. Finally, the selected descriptors underwent an outlier treatment process. For each descriptor, the first and third quartiles (Q1, Q3) were computed, and the interquartile range (IQR = Q3 - Q1) was calculated. Any records outside the range [Q1 - 1.5 * IQR, Q3 + 1.5 * IQR] were flagged as outliers and replaced with the nearest permitted value (Mauri, A., 2020). This outlier treatment step helped to improve the robustness and reliability of the subsequent modeling. By implementing this feature selection and preprocessing pipeline, the developed models were able to focus on the most relevant, uncorrelated, and well-behaved 3D structural descriptors. This optimization of the input features ultimately enhanced the performance and interpretability of the biodegradability classification models.

4 Machine Learning (ML) Methods

Four standard ML models were investigated in combination with different 3D molecular representation groups: Gradient Boosting (GB) (Nahm, F. S., 2022), XGBoost (XGB) (Moosbauer et al., 2021), random forest (RF) (Natekin and Knoll, 2013) and support vector machine SVM (Obikee and Happiness, 2014). Values of the hyperparameters for each model that showed the best cross validated area under the receiver

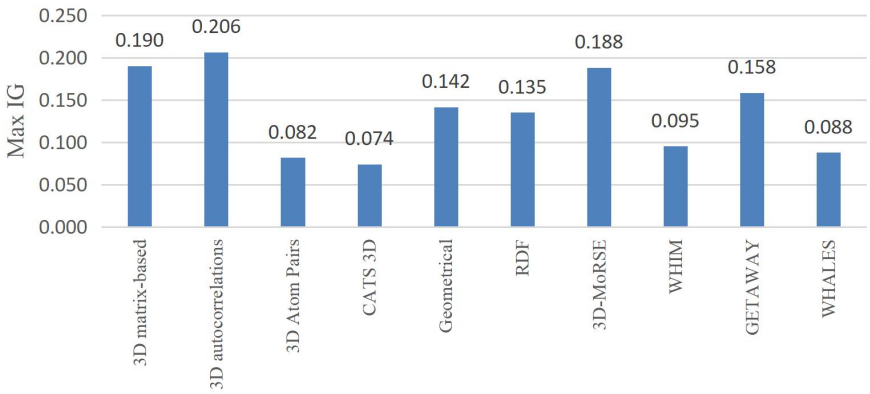


Figure 2. The maximum IG of each 3D molecular descriptor groups.

Table 1. Statistics of information gain (IG) of all descriptors for each 3-D molecular Descriptor Groups

Descriptor Group	Description	Count	Max IG	Mean IG	IG STD	Sum
3D Matrix-Based	Represent the 3D molecular structure using a matrix that encodes geometric relationships (distances, angles, torsions) between atom pairs.	103	0.190	0.048	0.048	4.989
3D Autocorrelations	Capture the 3D arrangement of functional groups and atoms by encoding the autocorrelation of 3D atomic properties (e.g., charge, electronegativity) across the molecular structure.	78	0.206	0.085	0.052	6.594
3D Atom Pairs	Provide a compact description of 3D molecular shape and potential interactions by representing the 3D distances between pairs of atoms.	31	0.082	0.013	0.021	0.388
CATS	Encode the topological distances between specific pharmacophore features within molecules, capturing spatial relationships in 3D space to aid in the analysis and prediction of molecular interactions and activities.	175	0.074	0.007	0.012	1.224
Geometrical	Represent the 3D shape of a molecule by capturing geometric properties like distances, angles, and torsions between atoms.	36	0.142	0.049	0.035	1.769
RDF	Describe the density of atoms at various distances from a central atom, providing a statistical representation of the 3D distribution of interatomic distances within a molecule.	210	0.135	0.017	0.020	3.597
3D-MoRSE	Derived from 3D atomic coordinates obtained from electron diffraction studies, these descriptors represent the 3D structure of molecules based on electron diffraction.	220	0.188	0.056	0.040	12.284
WHIM	Encode 3D shape, size, and atom distribution characteristics, providing an interpretable summary of key 3D molecular properties.	83	0.095	0.030	0.020	2.513
GETAWAY	Combine 3D geometric, topological, and atomic properties to represent a diverse set of 3D features relevant to molecular structure and function.	223	0.158	0.041	0.038	9.179
WHALES	Quantify the 3D distribution and properties of atoms within the molecular structure, potentially correlating with specific biological mechanisms.	33	0.088	0.038	0.031	1.238

operating characteristic curve (AUC) have been determined using Bayesian optimization (BO).

4.1 Gradient boosting (GB)

Gradient Boosting is an ensemble learning method that combines multiple decision trees to make predictions. It works by sequentially adding trees that minimize the errors of previous trees, using a gradient descent approach. This iterative process leads to highly accurate models, particularly for complex datasets. Key hyperparameters include the number of trees, the maximum depth of each tree, and the learning rate. Advantages of Gradient Boosting include its high accuracy, ability to handle diverse datasets, and robustness to outliers. However, it can be computationally expensive and susceptible to overfitting if not properly tuned.

4.2 XGBoost (XGB)

XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm for classification tasks. It excels at handling complex datasets and often outperforms other methods. XGBoost works by sequentially building an ensemble of decision trees, where each tree aims to correct the errors made by previous trees. It uses a gradient boosting algorithm to minimize the loss function at each iteration. Key hyperparameters include the number of trees, the maximum depth of each tree, the learning rate, and the fraction of data and features used for training each tree. XGBoost is known for its regularization techniques that prevent overfitting, leading to robust and accurate models. It is widely used in various domains, including finance, healthcare, and natural language processing.

4.3 Random forest (RF)

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It builds each tree on a random subset of the data and features, reducing variance and improving generalization. This ensemble approach often leads to high accuracy and robustness. Key hyperparameters include the number of trees, the maximum depth of each tree, and the number of features considered at each split. Advantages of Random Forest include its ability to handle high-dimensional data, its resistance to overfitting, and its relative ease of interpretation. However, it can be computationally expensive, especially for large datasets, and may not provide as much interpretability as simpler models.

4.4 Support vector machine (SVM)

Support Vector Machines (SVM) with a Radial Basis Function (RBF) kernel are powerful classification models that aim to find the optimal hyperplane to separate different classes in a high-dimensional feature space. The RBF kernel allows for non-linear decision boundaries, making SVM well-suited for complex datasets. Key hyperparameters include the regularization parameter (C) and the gamma parameter (γ) which controls the influence of individual data points. Advantages of RBF-SVM include high accuracy, robustness to outliers, and ability to handle non-linear relationships. However, it can be computationally expensive, especially for large datasets, and can be sensitive to the choice of hyperparameters.

4.5 Bayesian optimization

To optimize the hyperparameters of our four classification models (GB, XGB, RF, and SVM), we employed Bayesian optimization within the KNIME platform. This approach utilizes the Tree-structured Parzen Estimator (TPE) algorithm for efficient hyperparameter tuning (Penso et al., 2021). The Bayesian optimization process involves a two-phase approach:

•Warm-up Phase: An initial set of random hyperparameter combinations are sampled and evaluated. These evaluations are used to build a probabilistic model of the objective function, which in our case is the cross-validated Area Under the Curve (AUC).

•Exploration and Exploitation Phase: The TPE algorithm intelligently selects promising hyperparameter combinations based on the probabilistic model. This phase iteratively explores the search space, aiming to identify optimal hyperparameters that maximize the cross-validated AUC.

The search space for each hyperparameter is defined by specifying start and stop values, optionally restricting the search with a step size. The optimization loop runs for a predetermined number of iterations, sampling hyperparameter combinations with replacement. Bayesian optimization intelligently explores the search space, focusing on promising regions, and avoiding unnecessary evaluations. The probabilistic model within Bayesian optimization helps to handle complex optimization landscapes, making it more resilient to local optima. By identifying optimal hyperparameters, Bayesian optimization improves the performance of classification models, ensuring they are optimally tuned for specific tasks, such as biodegradability classification, thereby maximizing predictive accuracy and robustness.

5 Model/Group Evaluation

Five performance metrics have been used to compare the results of the optimized models GB, XGB, RF, and SVM with different 3D molecular descriptor groups. They include accuracy, sensitivity, specificity, F1 score, Kappa, and AUC. Most of them are computed using confusion matrix parameters like true positive TP, true negative (TN), false positive (FP), false negative (FN). The evaluation metrics are calculated as follows:

While simple accuracy measures, like the percentage of correct predictions, can be misleading when chance agreement is high, Kappa provides a more nuanced assessment

$$SEN = 100 \times \frac{TP}{TP+FN}, \quad (1)$$

$$SPEC = 100 \times \frac{TN}{TN+FP}, \quad (2)$$

$$ACC = 100 \times \frac{TP+TN}{TP+TN+FP+FN}, \quad (3)$$

$$F1_{score} = 100 \times \frac{2(SEN \times SPEC)}{SEN + SPEC}, \quad (4)$$

$$\text{Kappa} = \frac{(\text{Po} - \text{Pe})}{(1 - \text{Pe})}, \quad (5)$$

of agreement. It quantifies how much better the observed agreement is compared to what we'd expect by chance alone. Kappa is calculated by comparing the observed agreement (Po), the proportion of cases where both raters (or the model and true labels) agree, to the expected agreement by chance (Pe), which is the probability of agreement based on the marginal frequencies of each category. The Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve is a valuable metric for evaluating classification models (Pires et al., 2022). The ROC curve plots the model's ability to distinguish between classes at various thresholds, showing the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity). A perfect model would have a curve reaching the upper left corner (100% sensitivity and specificity), indicating perfect separation of classes. The AUC-ROC curve summarizes this performance, reflecting the model's overall ability to distinguish between positive and negative cases. A higher AUC indicates a stronger model, better at classifying readily biodegradable (RB) and non-readily biodegradable (NRB) molecules.

Bayesian optimization was employed to tune the hyperparameters of each machine learning model for each of the ten 3D molecular descriptor groups. The model/group combination that achieved the highest cross-validated AUC was selected for constructing the final model. Figure 3 illustrates the cross-validated AUC scores for the four models (SVM, XGBoost, Gradient Boosting, Random Forest) across the ten 3D descriptor groups.

Notably, the 3D MoRSE group achieved (0.91, 0.91, 0.91, and 0.9) AUC when using GB, XGB, RF and SVM respectively, while the 3D Autocorrelation group achieved an AUC of 0.90 with all models. The GETAWAY group also performed well, achieving AUC scores of 0.89-0.90 across the models. In contrast, the 3D Atom pairs, WHALES, and CATS 3D groups consistently yielded lower AUC scores across all models, indicating less effective performance in predicting readily biodegradability.

Table 2 summarizes all other performance metrics on the training and test datasets. These metrics include accuracy (ACC), sensitivity (Sen), specificity (Spec), F1 score (F1), Kappa, and AUC. The table shows that for the training subset, as expected, the three tree-based ensembles outperform the SVM model. The three ensembled achieved almost 100% training performance with all groups except 3D Atom Pairs and CATS 3D groups.

For the test data, SVM with 3D autocorrelation groups achieved the best performance with accuracy of 0.88, sensitivity of 0.83, specificity of 0.91, F1 score of, 0.82, Kappa of 0.74, and finally the AUC of0.93. These results include the best accuracy, sensitivity, F1, Kappa, and AUC compared to all other model/group combinations.

Based on the test dataset, the SVM model using the 3D Autocorrelation descriptor group achieved the best overall performance, with an accuracy of 0.88, sensitivity of 0.83, specificity of 0.91, F1 score of 0.82, Kappa of 0.74, and AUC of 0.93. This combination outperformed all other model/descriptor group combinations across these metrics. The Gradient Boosting (GB) model with the GETAWAY descriptor

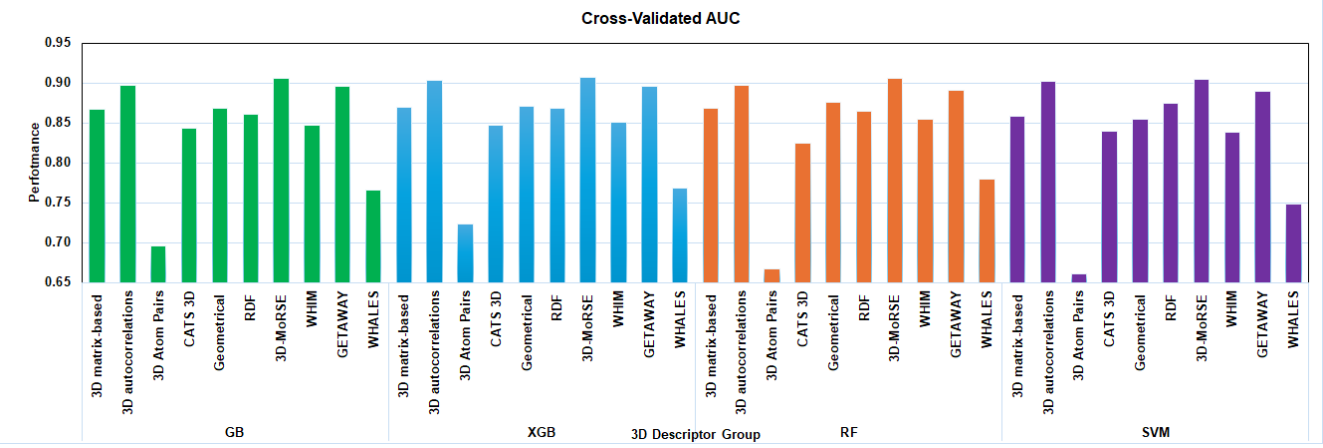


Figure 3. Cross-validated AUC of four ML models evaluated with ten 3D molecular descriptor groups.

group consistently demonstrated strong performance, ranking second to the SVM/3D Autocorrelation combination for most metrics.

•**Accuracy**: The GB/ GETAWAY model achieved an accuracy of 0.87, trailing slightly behind the top performer.

•**Sensitivity**: Similarly, the GB/GETAWAY combination achieved a sensitivity of 0.82, ranking second.

•**Specificity**: The XGBoost and GB models with 3D Autocorrelation descriptors tied for the highest specificity at 0.93.

•**F1 Score**: The GB/ GETAWAY model achieved an F1 score of 0.8, again ranking second.

•**Kappa**: The GB/ GETAWAY model achieved a Kappa score of 0.71, placing second. Importantly, no other model/descriptor combination achieved a Kappa score above 0.7.

•**AUC**: The GETAWAY group paired with XGBoost, GB, and RF achieved an AUC score of 0.93, mirroring the performance of the SVM/3D Autocorrelation combination.

These findings highlight the strong performance of both the SVM/3D Autocorrelation and GB/GETAWAY combinations. While SVM/3D Autocorrelation exhibited slightly superior overall performance, the GB model with GETAWAY demonstrated consistently high scores across a range of metrics, solidifying its position as a strong contender for biodegradability prediction. To comprehensively assess the performance of different 3D molecular descriptor sets we pooled the test metric values across all combinations. Figure 3 analyzes the performance of each descriptor group across all models. This figure demonstrates that the GETAWAY, 3D Autocorrelation, and 3D MoRSE descriptor groups consistently achieved the highest scores, indicating their effectiveness in predicting readily biodegradability. Conversely, the CATS 3D, WHALES, and 3D Atom Pairs groups performed the worst, suggesting they were less informative for this task. The remaining descriptor groups exhibited average performance.

The GETAWAY, 3D Autocorrelation, and 3D MoRSE descriptors consistently outperformed other groups. In contrast,

Table 2. Illustrates the contaminant factor for Al Shiqqah mining

Model	3D Descriptor	Training					Test					
		Acc	Sen	Spec	F1.00	KAPPA	Acc	Sen	Spec	F1	KAPPA	AUC
GB	3D matrix-based	1.00	1.00	1.00	1.00	1.00	0.82	0.77	0.84	0.73	0.60	0.86
	3D	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.87**	**0.72**	**0.93**	**0.78**	**0.68**	**0.93**
	3D Atom Pairs	0.84	0.52	0.99	0.67	0.58	0.70	0.28	0.90	0.38	0.21	0.69
	CATS 3D	0.98	0.97	0.99	0.98	0.96	0.77	0.63	0.84	0.64	0.47	0.82
	Geometrical	1.00	1.00	1.00	1.00	1.00	0.81	0.71	0.86	0.71	0.57	0.89
	RDF	1.00	1.00	1.00	1.00	1.00	0.84	0.77	0.87	0.76	0.64	0.90
	3D-MoRSE	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.85**	**0.73**	**0.90**	**0.76**	**0.65**	**0.91**
	WHIM	1.00	1.00	1.00	1.00	1.00	0.80	0.68	0.86	0.69	0.55	0.89
	GETAWAY	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.87**	**0.82**	**0.90**	**0.80**	**0.71**	**0.93**
	WHALES	1.00	1.00	1.00	1.00	1.00	0.72	0.52	0.81	0.54	0.34	0.78
XGB	3D matrix-based	1.00	1.00	1.00	1.00	1.00	0.79	0.73	0.82	0.70	0.54	0.86
	3D	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.86**	**0.70**	**0.93**	**0.76**	**0.65**	**0.92**
	3D Atom Pairs	0.76	0.45	0.91	0.55	0.39	0.71	0.33	0.89	0.42	0.25	0.70
	CATS 3D	0.96	0.92	0.98	0.94	0.91	0.80	0.67	0.87	0.69	0.55	0.86
	Geometrical	0.99	0.97	0.99	0.98	0.97	0.84	0.73	0.89	0.74	0.62	0.87
	RDF	1.00	1.00	1.00	1.00	0.99	0.84	0.81	0.86	0.77	0.64	0.90
	3D-MoRSE	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.83**	**0.71**	**0.88**	**0.72**	**0.60**	**0.91**
	WHIM	0.97	0.93	0.98	0.94	0.92	0.83	0.72	0.89	0.74	0.62	0.89
	GETAWAY	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.86**	**0.80**	**0.88**	**0.78**	**0.67**	**0.93**
	WHALES	1.00	1.00	1.00	1.00	1.00	0.71	0.50	0.82	0.53	0.32	0.76
RF	3D matrix-based	1.00	1.00	1.00	1.00	1.00	0.82	0.76	0.84	0.73	0.59	0.88
	3D	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.82**	**0.68**	**0.89**	**0.71**	**0.58**	**0.92**
	3D Atom Pairs	0.68	0	1.00	-	0	0.68	0	1.00	-	0	0.61
	CATS 3D	0.94	0.83	0.99	0.89	0.85	0.78	0.44	0.93	0.56	0.42	0.83
	Geometrical	1.00	1.00	1.00	1.00	1.00	0.84	0.71	0.91	0.74	0.63	0.89
	RDF	1.00	1.00	1.00	1.00	1.00	0.83	0.73	0.87	0.73	0.60	0.90
	3D-MoRSE	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.85**	**0.72**	**0.91**	**0.76**	**0.65**	**0.92**
	WHIM	1.00	1.00	1.00	1.00	1.00	0.81	0.66	0.89	0.70	0.56	0.90
	GETAWAY	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.86**	**0.78**	**0.89**	**0.78**	**0.67**	**0.93**
	WHALES	1.00	1.00	1.00	1.00	1.00	0.74	0.49	0.86	0.55	0.37	0.80
SVM	3D matrix-based	0.82	0.66	0.89	0.70	0.57	0.80	0.66	0.86	0.68	0.53	0.86
	3D	**0.95**	**0.91**	**0.97**	**0.92**	**0.88**	**0.88**	**0.83**	**0.91**	**0.82**	**0.74**	**0.93**
	3D Atom Pairs	0.69	0.14	0.95	0.23	0.11	0.71	0.16	0.97	0.26	0.16	0.69
	CATS 3D	0.91	0.85	0.93	0.85	0.78	0.81	0.73	0.84	0.71.00	0.57	0.87
	Geometrical	0.86	0.75	0.91	0.77	0.67	0.82	0.72	0.87	0.72	0.59	0.87
	RDF	0.93	0.89	0.94	0.89	0.83	0.82	0.79	0.83	0.74	0.60	0.91
	3D-MoRSE	**0.92**	**0.88**	**0.94**	**0.88**	**0.82**	**0.85**	**0.8**	**0.88**	**0.78**	**0.67**	**0.91**
	WHIM	0.90	0.82	0.93	0.83	0.76	0.84	0.76	0.88	0.75	0.64	0.91
	GETAWAY	**0.89**	**0.81**	**0.93**	**0.83**	**0.75**	**0.85**	**0.78**	**0.88**	**0.77**	**0.65**	**0.92**
	WHALES	0.84	0.57	0.97	0.69	0.59	0.73	0.46	0.86	0.52	0.34	0.73

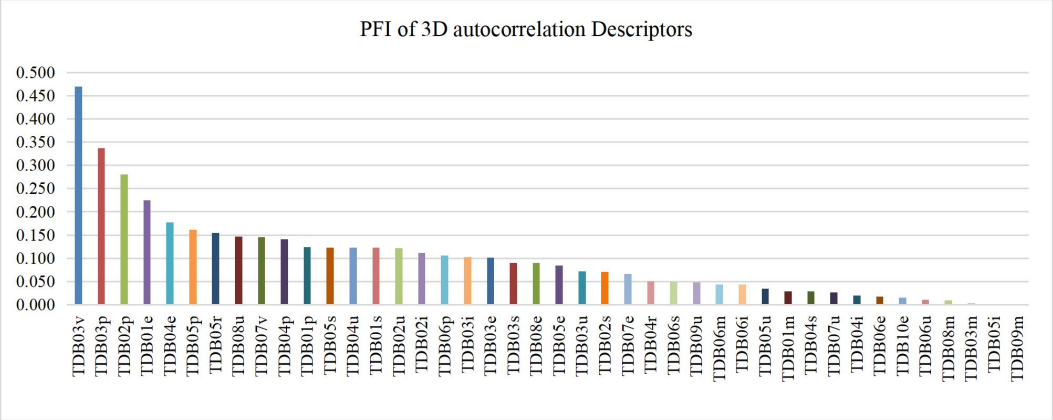


Figure 4. PFI importance analysis of the 3D autocorrelation molecular descriptors (42 predictive descriptors) with SVM model. Number of permutations (10) and performance metric Cohen's Kappa.

the 3D Atom Pairs, CATS 3D, and WHALES descriptors achieved significantly lower performance. The performance differences between the descriptor groups likely stem from the specific types of structural information they capture and their relevance to readily biodegradability.

•**GETAWAY**: This descriptor group captures a broad range of 3D features, including geometric, topological, and atomic properties. It excels at capturing the overall shape, electrostatic potential, and surface characteristics of a molecule, all of which are crucial for interactions with enzymes and microorganisms involved in biodegradation.

•**3D Autocorrelation**: These descriptors encode the spatial distribution of specific atom types or functional groups. By capturing the 3D arrangement of these features, they can identify crucial structural motifs associated with biodegradation pathways.

•**3D MoRSE**: These descriptors represent a molecule's 3D structure based on electron diffraction studies. They effectively capture a wide range of molecular properties, including size, shape, and electronic effects, which can significantly impact a molecule's susceptibility to biodegradation.

Other descriptor groups (especially 3 Atom Pairs) might overlook critical long-range interactions and the overall molecular shape, both of which are significant for biodegradability. Additionally, they might not directly capture the intricate mechanisms of biodegradation, which involve a complex interplay of factors.

6 Feature Importance

Given its superior performance, the combination SVM/3D autocorrelation was selected for further investigation using the Permutation Feature Importance (PFI) (Qu et al., 2023), Shapley analysis (SHAP) (Ramraj et al., 2016), and Partial Dependence Plot (PDP) algorithms (Ramraj et al., 2016) and (Rocha and Sheen, 2016). PFI assesses the importance of individual features by measuring the impact of randomly shuffling their values on the model's performance (Singh et al., 2021). It calculates the difference between the model's score using all original features and the score achieved when one specific feature is randomly permuted. The process is repeated multiple times for each feature, and the average difference in scores, along with the standard deviation, is calculated. A larger decrease in performance after shuffling a feature indicates that it is more important to the model's predictions. Figure 4 shows PFI computed successfully using Cohen's kappa and the class of interest "RB" for SVM and 3D autocorrelation descriptors.

PFI analysis provided insights into the relative importance of all predictive features in the SVM model. The results of the analyses showed that the first five descriptors (TDB03v, TDB03p, TDB02p, TDB01e, and TDB04e) were relatively more important features in the SVM prediction model. This information is valuable for understanding the structural features that influence biodegradability and for developing more accurate and robust prediction models. The 3D autocorrelation descriptors like TDB capture the spatial distribution of certain molecular properties within a molecule. They are essentially measures of how different chemical features are correlated in 3D space. The fact that these descriptors are most important suggests that the following aspects of the molecule likely play a crucial role in biodegradability.

•**TDB03v**: Measures the autocorrelation of the presence of atoms with a specific property at a distance of 3 bonds. This might indicate that the presence of specific functional groups or atom types spaced 3 bonds apart influences biodegradability.

•**TDB03p**, **TDB02p**: These descriptors relate to the autocorrelation of properties over specific distances, suggesting that the arrangement of atoms and functional groups within a certain range is critical.

•**TDB01e**, **TDB04e**: These descriptors emphasize the importance of electrostatic properties, particularly at specific distances within the molecule.

Understanding the importance of these descriptors allows us to establish relationships between molecular structure and

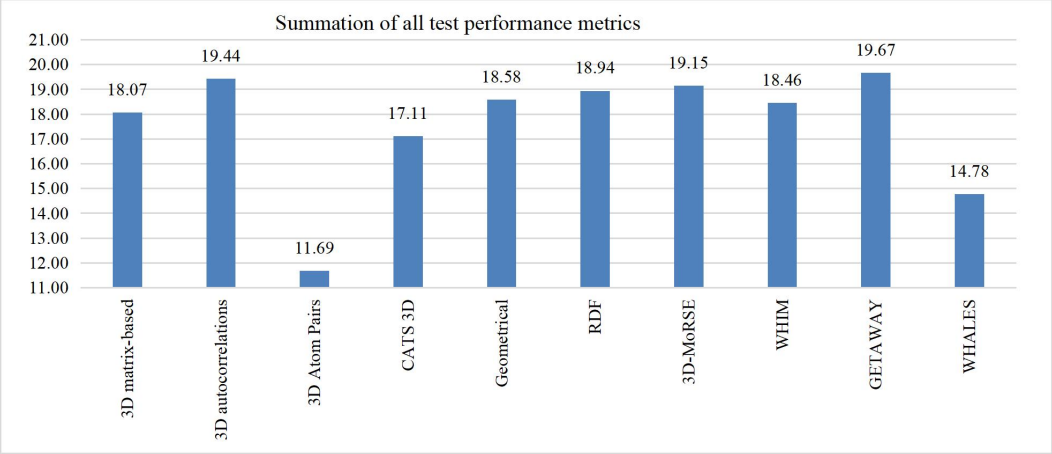


Figure 5. Summation of all performance metrics of 3D molecular descriptor groups across all ML model.

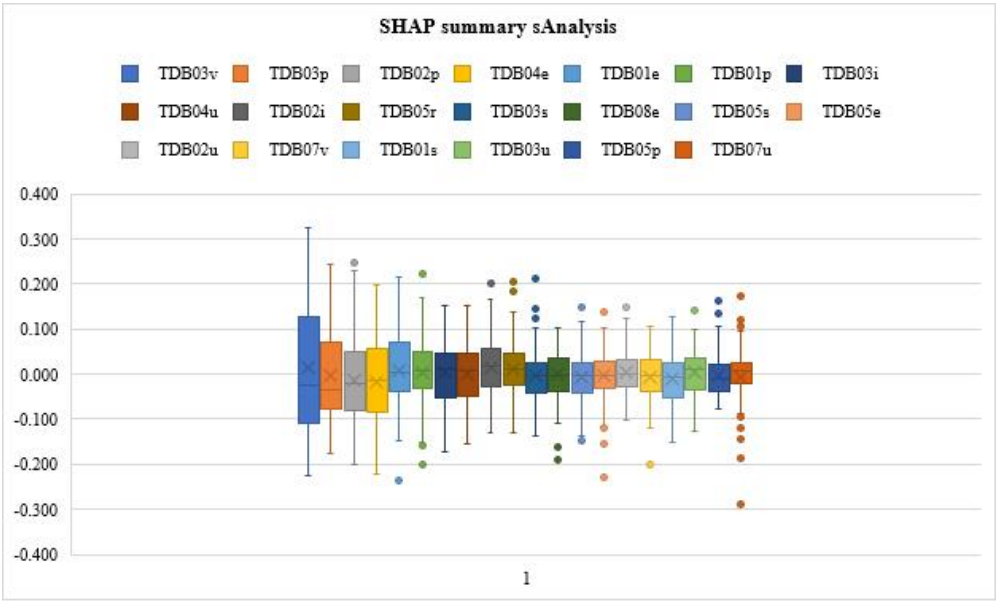
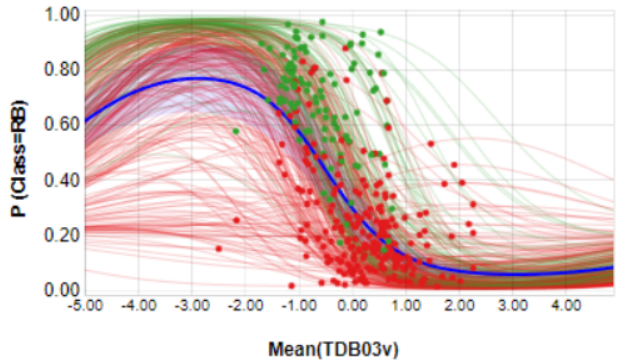


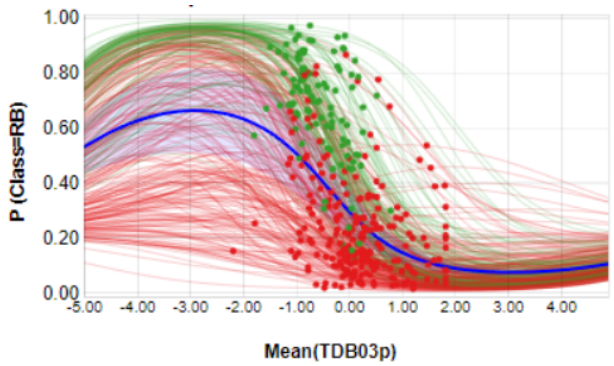
Figure 6. SHAP summary plot of the top 20 features and class RB for 100 test records.

biodegradability. This can help you design molecules with desired biodegradability properties. To further validate the importance of the identified 3D autocorrelation descriptors, we employed the SHAP (SHapley Additive exPlanations) method. SHAP is generally used by researchers to solve black-box problems associated with ML models and to assess the interaction or synergy of two variables. We analyzed 100 test records to assess the relationship between each descriptor and the predicted RB class (Figure 6). The SHAP summary plot ranks the descriptors based on the mean absolute SHAP value, revealing the top 20 most influential descriptors. A positive SHAP value indicates a positive correlation between the descriptor and the RB class, with larger values signifying a greater contribution to the prediction. The SHAP analysis confirmed the findings of the Permutation Feature Importance (PFI) analysis, identifying the same top descriptors: TDB03v, TDB03p, TDB02p, TDB04e, and TDB01e. This consistency across two independent methods strengthens our confidence in the predictive importance of these specific 3D autocorrelation descriptors for biodegradability classification.

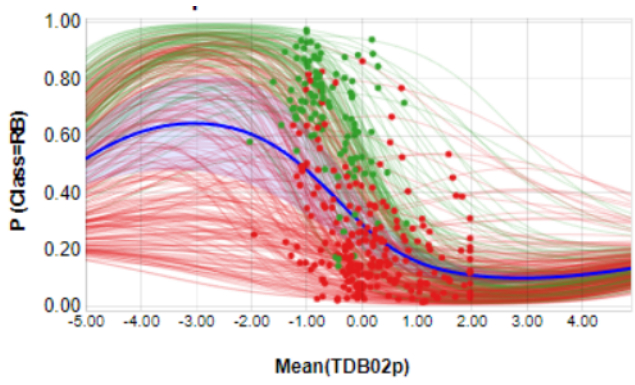
To further elucidate the relationship between the top 3D autocorrelation descriptors and predicted biodegradability, we employed Partial Dependence Plot (PDP) analysis. PDP analysis provides a graphical representation of how changes in individual descriptor values influence the predicted probability of a compound being readily biodegradable (RB). Figure 7 illustrates the PDPs for the five most influential 3D autocorrelation descriptors (TDB03v, TDB03p, TDB02p, TDB01e, and TDB04e) identified in our SVM model. Interestingly, four of these descriptors (TDB03v, TDB03p, TDB02p, and TDB01e) displayed a consistent trend: as their values increased, the predicted probability of RB decreased (6a), (6b), (6c), and (6d). This suggests that higher values for these descriptors are associated with lower biodegradability. In



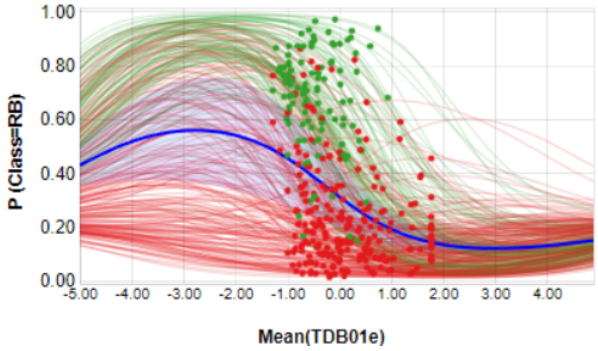
(a)



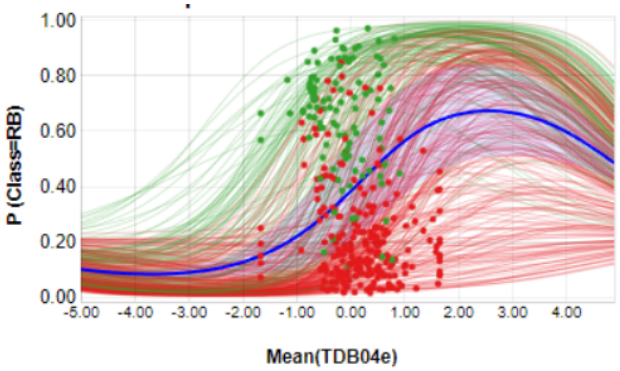
(b)



(c)



(d)



(e)

contrast, the descriptor TDB04e exhibited an inverse relationship, where increasing values corresponded to a higher predicted RB probability (7e). This detailed PDP analysis provides valuable insights into the complex relationships between specific 3D structural features and biodegradability, enhancing our understanding of the model's predictive behavior and highlighting potentially important structural motifs for biodegradability.

7 Conclusion

This research provides critical insights into the application of 3D molecular descriptors for predicting chemical biodegradability, a crucial factor in advancing sustainable chemical manufacturing and achieving the goals of SDG 11 (Sustainable Cities and Communities) and SDG 12 (Responsible Consumption and Production). Our comprehensive investigation of ten descriptor groups and machine learning algorithms revealed that GETAWAY, 3D Autocorrelation, and 3D MoRSE descriptors consistently demonstrated superior performance in predicting readily biodegradability. This underscores the importance of capturing detailed 3D structural information, particularly the spatial distribution of functional groups and atomic properties, for accurate biodegradability assessment.

The superior performance of 3D autocorrelation descriptors is attributed to their ability to capture the spatial arrangement of functional groups and atoms, long-range interactions, and key structural motifs. This descriptor group combines 3D geometric, topological, and atomic properties, offering a comprehensive representation of a molecule's structure. This comprehensive approach likely allows GETAWAY to capture diverse aspects of a molecule's interaction with biodegradation pathways, leading to improved prediction accuracy. The findings of this study have significant implications for the design of sustainable chemical products and processes. By utilizing 3D molecular descriptors and machine learning algorithms to predict biodegradability, we can:

Contribute to building sustainable cities: By promoting the development and use of readily biodegradable chemicals, we can minimize the environmental burden associated with chemical manufacturing and waste disposal, leading

to cleaner and healthier urban environments. Advance responsible consumption and production: By identifying and prioritizing the use of biodegradable chemicals, we can move towards a circular economy model that reduces resource depletion and minimizes environmental pollution. The high accuracy of the SVM model trained on 3D autocorrelation descriptors (0.88 accuracy, 0.83 sensitivity, 0.91 specificity, 0.82 F1-score, 0.74 Cohen's Kappa statistic, and 0.93 AUC) provides further evidence for the potential of this approach. By leveraging 3D molecular descriptors and machine learning, we can develop more sustainable chemical products and processes, fostering a healthier and more sustainable future aligned with the objectives of SDG 11 and SDG 12.

Author Contributionts

A.M. Elsayad was responsible for conceptualization, methodology, writing the original draft, and supervision. H.Y. Ahmed handled data curation, formal analysis, and writing - review and editing. K.A. Elsayad conducted investigation, software development, and visualization. A. N. Khan contributed resources, validation, and writing - review and editing. A.E.B Hassan managed project administration, funding acquisition, and writing - review and editing. M.M. H Mustafa performed data curation and formal analysis. A.A. Ali worked on software development and visualization. S.A. Mukhtar provided resources and validation.

Conflict of Interest

The authors declare no conflicts of interest.

Acknowledgement

The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU-2023/SDG/28).

References

Abdullah, D. M. and Abdulazeez, A. M., 2021. Machine learning applications based on SVM classification a review. Qubahan Academic Journal., 1(2): 81-90. https://doi.org/10.1080/17425255.2023.2294939

Ahmadi, S., Ketabi, S., Javan. M. J., 2023. Molecular Descriptors in QSPR/QSAR Modeling. In QSPR/QSAR Analysis Using SMILES and Quasi-SMILES., pp. 25-56. Cham: Springer International Publishing. https://doi.org/10.1007/s10462-020-09896-5

Bahia, M. S., Kaspi, O., Touitou, M., Binayev, I., 2023. A comparison between 2D and 3D descriptors in QSAR modeling based on bio-active conformations. Molecular Informatics ., 42(4): 2200186. https://doi.org/10.1039/D2DD00099G

Consonni, V. and Todeschini, R., 2010. Molecular descriptors. Recent advances in QSAR studies: methods and applications: 29-102. https://doi.org/10.6026/97320630013154

Chinedu, E., Arome, D., Ameh, F. S., 2013. A new method for determining acute toxicity in animal models. Toxicol. Int., 20(3): 224-226. https://doi.org/10.4103/0971-6580.121674

Haddouchi, M. and Berrado, A., 2019. A survey of methods and tools used for interpreting random forest. In 2019 1st International Conference on Smart Systems and Data Science (ICSSD), USA, 1-6. https://doi.org/10.1109/ICSSD47982.2019.9002770

Lombardo, A., Pizzo, F., Benfenati, E., 2014. A new in silico classification model for ready biodegradability, based on molecular fragments. Chemosphere 10(8): 10-16. https://doi.org/10.1016/j.chemosphere.2014.02.073

Lu, S. C., Swisher, S. L., Chung, C., 2023. On the importance of interpretable machine learning predictions to inform clinical decision making in oncology. Frontiers in Oncology 13: 1129380. https://doi.org/10.3389/fonc.2023.1129380

Mansouri, K., Ringsted, T., Ballabio, D., 2013. Quantitative structure–activity relationship models for ready biodegradability of chemicals. Journal of chemical information and modeling 53(4): 867-878. https://doi.org/10.1021/ci4000213

Mauri, A. and Bertola, M., 2022. Alvascience: A new software suite for the QSAR workflow applied to the blood–brain barrier permeability. International Journal of Molecular Sciences., 23(21): 12882. https://doi.org/10.3390/ijms232112882

Mauri, A., 2020. alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints. Ecotoxicological QSARs., 1(32): 801-820. https://doi.org/10.1007/978-1-0716-0150-1_32

Moosbauer, Julia, Julia Herbinger, Giuseppe Casalicchio., 2021. Explaining hyperparameter optimization via partial dependence plots. Advances in Neural Information Processing Systems., 34(2021): 2280-2291. https://doi.org/10.5555/3540261.3540436

Natekin, A. and Knoll, A., 2013. Gradient boosting machines, a tutorial. Frontiers in neurorobotics 7: 1-21. https://doi.org/10.3389/fnbot.2013.00021

Nahm, F. S., 2022. Receiver operating characteristic curve: overview and practical use for clinicians. Korean journal of anesthesiology., 75(1) : 25-36. https://doi.org/10.4097/kja.21209

Obikee, A. C., Godday, U. E. and Happiness O. Obiora-Ilouno, 2014. Comparison of outlier techniques based on simulated data. Open Journal of Statistics., 4(7): 536-56. https://doi.org/10.4236/ojs.2014.47051

Penso, M., Pepi, M., Fusini, L., 2021. Predicting long-term mortality in TAVI patients using machine learning techniques. Journal of Cardiovascular Development and Disease., 8(4): 1-14. https://doi.org/10.3390/jcdd8040044

Pires, J. R. A., Souza, V. G. L. S, Fuciños, P., Pastrana, L., Fernando, A. L., 2022. Methodologies to assess the biodegradability of bio-based polymers—current knowledge and existing gaps. Polymers 14(7): 1359. https://doi.org/10.3390/polym14071359

Qu, K., Xu, J., Hou, Q., Qu, K., Sun, Y., 2023. Feature selection using Information Gain and decision information in neighborhood decision system. Applied Soft Computing., 136: 110100. https:/doi.org//10.1016/j.asoc.2023.110100

Ramraj, S., Uzir, N., Sunil, R., Banerjee, S., 2016. Experimenting XGBoost algorithm for prediction and classification of different datasets. International Journal of Control Theory and Applications., 9(40): 651-662.

Rigatti, S. J., 2017. Random forest. Journal of Insurance Medicine., 47(1): 31-39. https://doi.org/10.17849/insm-47-01-31-39.1

Rocha, W. F. C. and Sheen, D. A., 2016. Classification of biodegradable materials using QSAR modelling with uncertainty estimation. SAR and QSAR in Environmental Research., 27(10): 799-811. https://doi.org/10.1080/1062936X.2016.1238010

Singh, A. K., Bilal, M., Iqbal, H. M. N., 2021. Trends in predictive biodegradation for sustainable mitigation of environmental pollutants: Recent progress and future outlook. Science of The Total Environment., 770: 144561. https://doi.org/10.1016/j.scitotenv.2020.144561

Wang, X., Jin, Y., Schmitt, S., Olhofer. M., 2023. Recent advances in Bayesian optimization. ACM Computing Surveys., 55(13): 1-36. https://doi.org/10.1145/3582078