PlastiGuard: ML/Molecular Descriptor based Biodegradability Analyzer of Persistent Chemical Carcinogens in Plastics

Riya Bharath

Abstract

Plastics pose a significant threat to human health by introducing carcinogenic chemicals through inhalation and digestion. Carcinogens such as Bisphenol A (BPA), Diethyl Phthalate, Per- and Polyfluoroalkyl Substances (PFAS), and Polyvinyl Chloride (PVC) disrupt reproduction, growth, and metabolism by mimicking hormonal activity. Their persistence in the environment increases human exposure, particularly when non-biodegradable.

To evaluate biodegradability, four machine learning models—logistic regression, k-nearest neighbors, random forest, and neural network—were trained on molecular descriptor data generated using RDKit. The dataset was refined through Principal Component Analysis (PCA) to assess its impact on performance. Results indicate that the neural network achieved the highest accuracy with RDKit descriptor data (88.1%) and PCA-transformed data (78.2%). Across all models, the RDKit dataset outperformed PCA-applied data, with accuracy improvements of +9.9% for the neural network, +14.3% for logistic regression, +9.7% for random forest, and +6.9% for k-nearest neighbors.

Notably, all models classified BPA, Perflurorooctanic Acid (PFOA), and Perfluorooctanesulfonic Acid (PFOS) as non-biodegradable, identifying them as the most persistent and hazardous carcinogens in plastics. To identify safer alternatives, cosine similarity was employed to pinpoint structurally similar compounds, which were subsequently assessed for biodegradability. This approach revealed that viable naturally produced substitutes include 3-Propionylchromone, N-Acetyl-D-glucosamine, and Butyl Paraben. This study highlights the potential of machine learning in identifying biodegradable alternatives to persistent carcinogens in plastics. By reducing the environmental and health risks associated with plastic-related toxins,

these findings offer a crucial step toward more sustainable and responsible plastic manufacturing practices.

Introduction

Plastic degradation into microplastics poses a growing concern due to its widespread impact on human exposure to harmful chemicals. Microplastics enter the human body through inhalation and ingestion, and studies show that around 77% of people have microplastics in their blood. Plastics are made with carcinogenic compounds such as BPA, Diethyl Phthalate, PFAS, and PVC, which mimic hormones and disrupt reproduction, growth, and metabolism, ultimately increasing cancer risk.

Biodegradability, defined as the ability of a material to be decomposed by organisms like bacteria or fungi and reintegrated into nature, is critical in evaluating plastic safety. Non-biodegradable plastics are particularly harmful as they do not break down easily, leading to prolonged exposure to toxic additives. Currently, biodegradation tests take weeks to years to complete, severely limiting the ability to screen chemicals efficiently. As a result, efforts to identify persistent, high-risk plastic additives are significantly delayed, slowing progress in creating safer plastics.

To accelerate this process, this project explores a computational approach to predict chemical biodegradability using machine learning. We extracted molecular features using cheminformatics tools like RDKit, trained classification models, and incorporated similarity analysis to suggest safer alternatives to harmful compounds. This approach draws from prior research in computational toxicology and cheminformatics, offering a scalable pathway to support safer plastic design. In prior research, machine learning models like QSAR and

classification trees have been used to predict chemical toxicity and biodegradability, speeding up environmental assessments compared to lab tests. However, these models mostly target broad chemical groups such as pharmaceuticals or pesticides, not specifically carcinogenic plastic additives.

This project focuses on carcinogenic plastic additives, creating tailored models to predict their biodegradability. It also uniquely uses structural similarity analysis to suggest safer, biodegradable alternatives, emphasizing actionable substitution. By building on the foundational chemistry of these compounds and existing testing limitations, this work aims to fill a critical gap in identifying persistent plastic additives more efficiently.

Dataset

Overview

This study uses the QSAR Biodegradation Dataset from the UCI Machine Learning Repository which contains 1,055 chemical compounds classified as either biodegradable or non-biodegradable. Each compound is represented by 41 numerical molecular descriptors (Ex: molecular weight, number of heteroatoms, number of hydrogen bond donors/acceptors, etc.) These features describe a compound's structural and chemical properties.The model uses this dataset to train and test a binary classification algorithm that predicts biodegradability based on these features.

Data Cleaning and Preprocessing

To ensure consistency and maximize the use of available information about each compound, the process retained all 41 features. Instead of using the original descriptors from the dataset, we

regenerated them with RDKit to ensure consistency between the training and testing sets and removed all samples containing null values. Then, we encoded the categorical target variable of biodegradability using Python's LabelEncoder, assigning "1" to "Biodegradable (RB)" and "0" to "Non-biodegradable (NRB)."

Before training, we standardized all feature values using StandardScaler, which scales features to have a mean of zero. This step is essential for models like Logistic Regression and K-Nearest Neighbors, which are sensitive to the magnitude of input values. Without scaling, large-valued features could disproportionately influence model behavior.

We split the dataset into 75% for training and 25% for testing.

Limitations

The dataset exhibits a slight class imbalance, with 699 biodegradable and 356 non-biodegradable compounds. The evaluation process accounted for this imbalance, especially when analyzing precision, recall, and false positive rates.

Methodology

Overview

This research aimed to develop machine learning models that predict the biodegradability of chemical carcinogens using molecular structure data. The approach involved generating features from chemical structures via RDKit, reducing data dimensionality with Principal Component Analysis (PCA), training and evaluating multiple classifiers, and ultimately applying the models to real-world chemical pollutants to explore safer biodegradable alternatives.

Feature Generation and Preprocessing

We converted each chemical's SMILES string into a molecular object using RDKit, which then generated 41 molecular descriptors per compound. These descriptors served as the features for all models. As described in the Dataset section, the preprocessing pipeline encoded and prepared the dataset before training and testing.

Dimensionality Reduction with PCA

Before model training, the analysis applied Principal Component Analysis (PCA) to reduce data dimensionality while preserving most of its variance. PCA helps simplify complex datasets, reduce noise, and improve model performance and interpretability. We transformed both training and testing data sets and retained the top 7 principal components based on explained variance as input features for all models. Each machine learning model was tested with the original RDkit data and transformed PCA data independently.

Model Training and Evaluation

The modeling phase trained and evaluated four machine learning models using the PCA-reduced, standardized data and the original data: Logistic Regression (LR), K-Nearest Neighbors (KNN), Random Forest (RF), and a Neural Network (NN).

Logistic Regression

The training process fits the Logistic Regression model using optimized parameter(s) determined via the .best_params_ attribute in scikit-learn's tuning process.

As a linear model, Logistic Regression established a baseline for performance. The model learned from the scaled PCA-transformed training data and made predictions on the corresponding test data.

K-Nearest Neighbors

The KNN model classified test points by comparing them to the optimized parameter(s) determined via the .best_params_ attribute in scikit-learn's tuning process of 5 nearest neighbors in the training set using Euclidean distance.

Random Forest

The training process fit the Random Forest model using optimized parameters of max_depth=5 determined via the .best_params_ attribute in scikit-learn's tuning process:

By limiting tree depth, the model avoided overfitting while remaining interpretable. This ensemble approach captured complex relationships between molecular features and biodegradability.

Neural Network

The implementation used a Multi-Layer Perceptron (MLP) neural network with the following optimized parameter(s) determined via the .best_params_ attribute in scikit-learn's tuning process:

- hidden_layer_sizes=(100, 78)
- activation="relu"
- max_iter=300

The model achieved strong training accuracy, though its higher complexity occasionally caused overfitting, as seen in the performance gap between training and test scores.

Performance Metrics

The evaluation process measured each model's performance using several metrics to assess both accuracy and class balance:

- Accuracy Score
- Precision Score
- Recall (True Positive Rate)
- False Positive Rate
- Confusion Matrix
- ROC curve

Application to Real-World Carcinogens

After validating the models, we used them to classify several known environmental carcinogens by feeding in their SMILES representations and generating descriptors with RDKit:

- Bisphenol A (BPA)
- Perfluorooctanoic acid (PFOA)
- Polyvinyl chloride (PVC)
- Diethyl phthalate (DEP)
- Perfluorooctanesulfonic acid (PFOS)

Each chemical went through the trained models, which labeled them as biodegradable (1) or non-biodegradable (0).

Identifying Biodegradable Alternatives

To suggest safer replacements for harmful chemicals, the analysis performed a cosine similarity search. This process compared the vector representations (molecular descriptors) of the carcinogens to other compounds in the dataset. The goal was to identify chemicals that:

- Have similar molecular structures (high cosine similarity)
- Are classified as biodegradable (label = 1)

This similarity search enabled the discovery of structurally similar but eco-friendlier alternatives that could potentially replace persistent pollutants in consumer products or industrial processes.

Results and Discussion

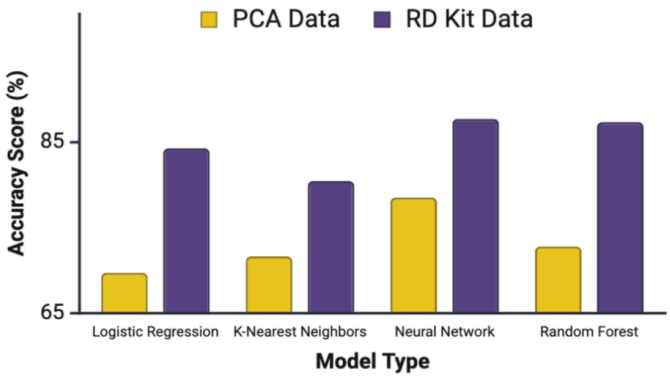

Fig. 1: Accuracy Score (%) of PCA vs. RD-Kit Data

While Principle Component Analysis (PCA) is typically used to reduce dimensionality of training data to improve model performance, it did not preserve the critical chemical characteristics needed for accurate predictions. As shown in Figure 1, all four models performed

better when trained with RD-Kit descriptors compared to PCA-based models. Consequently, the RD-Kit trained models were used for the remainder of the analysis.

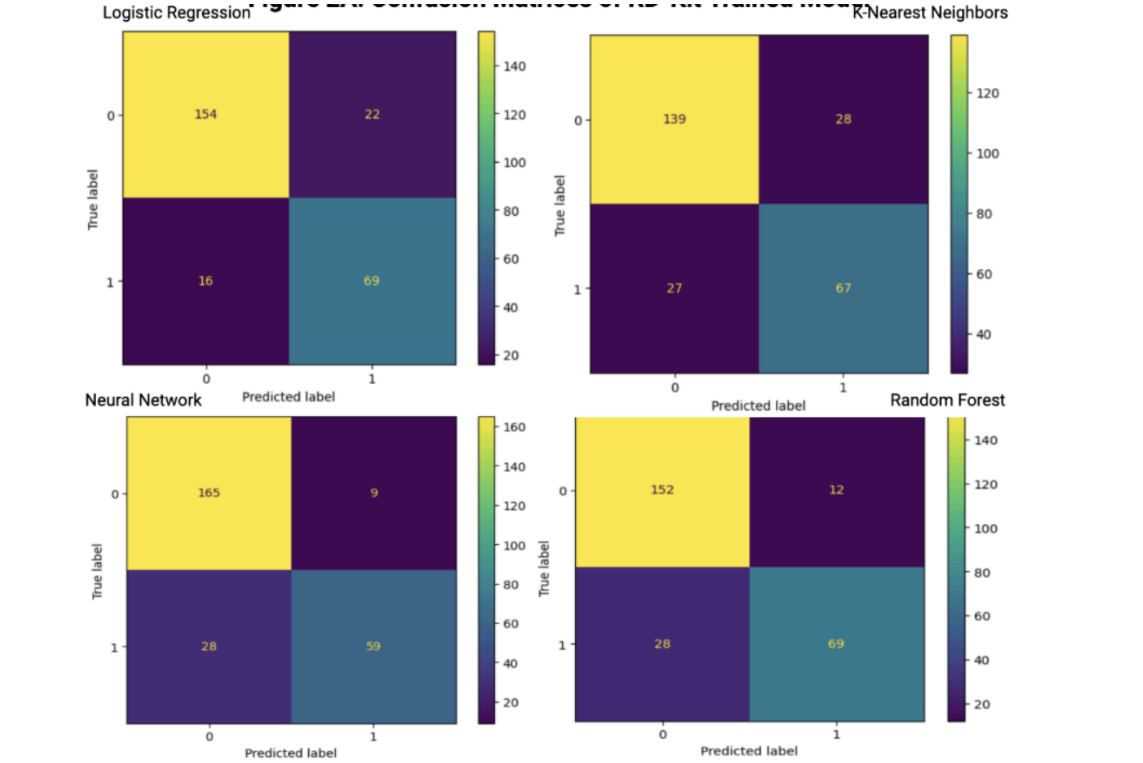

Fig. 2A: Confusion Matrices of RD-Kit Trained Model

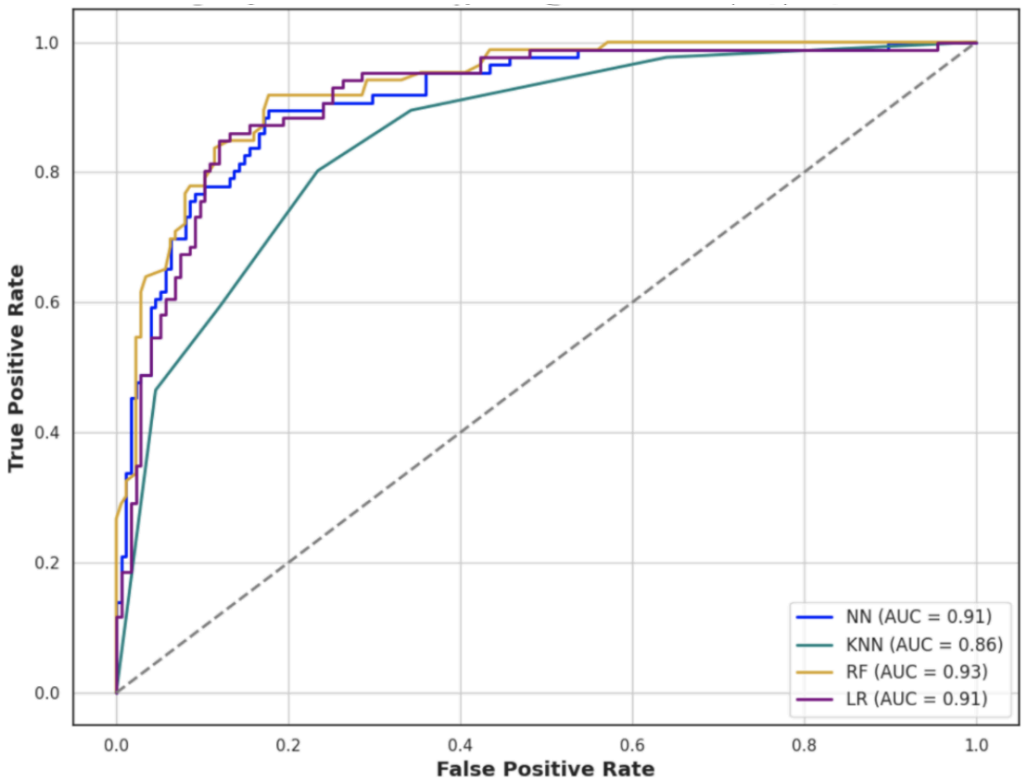

Fig. 2B: Receiver Operating Characteristic (ROC) Curve

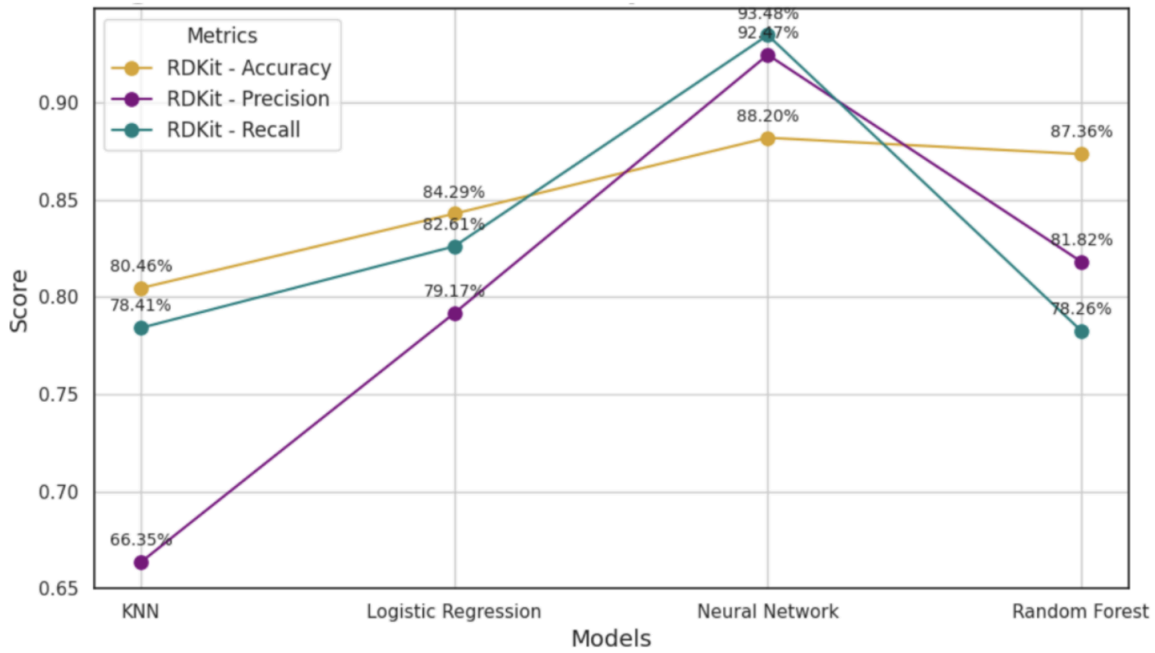

Fig. 2C: Performance Comparison of RD-Kit Trained Models

Among the models tested, the neural network and random forest algorithms demonstrated

the best overall performance. The neural network model achieved the highest accuracy of 88.20%, precision of 92.47%, and recall score of 93.48% (Figure 2A) and the lowest false positive rate of 13.24% (Figure 2C), indicating its superior ability to correctly classify biodegradable chemicals and minimize false negatives.

The random forest model, while slightly behind in accuracy of 87.36%, still performed strongly, with a precision of 81.82%, recall of 78.26% (Figure 2A), and false positive rate of 14.81% (Figure 2C) Receiver Operating Characteristic (ROC) curves further confirmed the random forest model's strength. A higher area under the curve (AUC) indicates a model's superior performance in distinguishing between biodegradable and non-biodegradable chemicals.The ROC curve, which graphically represents a model's ability to distinguish between classes, shows that the random forest model achieved the best AUC of 0.93 compared to the neural network's AUC of 0.91 (Figure 2B). In this study, the random forest model's ROC performance confirmed its robustness and effectiveness in classifying chemicals accurately. The neural network and random forest models generated the best results because they are able to trace more complicated relationships in data by creating new nonlinear feature representations.

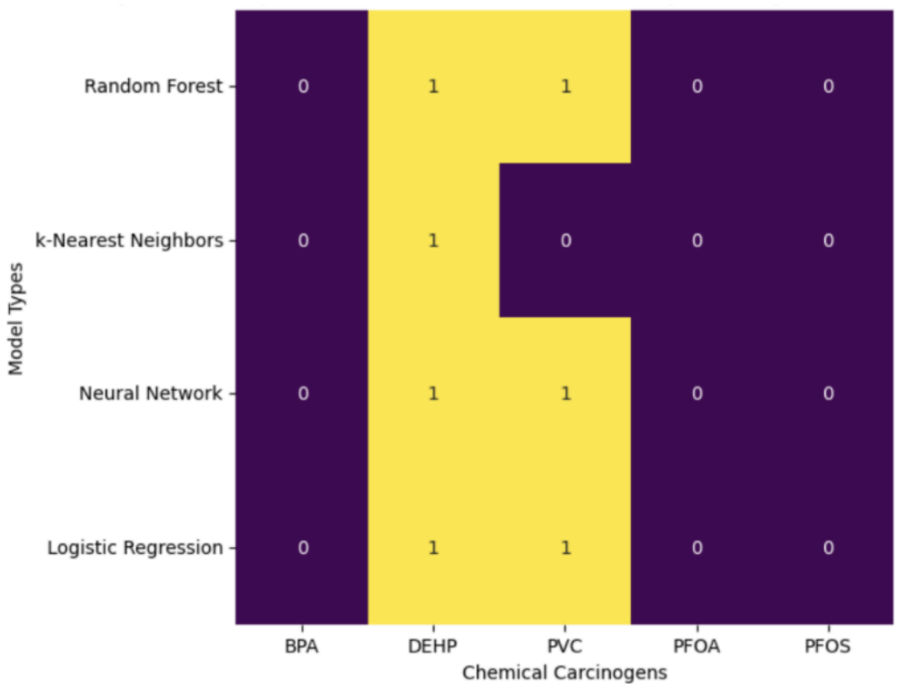

Fig. 3: Biodegradability Classification of Common Carcinogenic Chemicals in Plastics

0 - Non-Biodegradable

1 - Biodegradable

After establishing and evaluating the models, they were used to classify the top five most common carcinogenic chemicals used in plastic production: Bisphenol A (BPA), Diethyl Phthalate (DEHP), Polyvinyl Chloride (PVC), Perfluorooctanoic Acid (PFOA), and Perfluorooctanesulfonic Acid (PFOS). Among the five chemicals analyzed, DEP is the most biodegradable. It breaks down relatively quickly under both aerobic and anaerobic conditions, especially in wastewater environments. BPA is less biodegradable; while some microbes can degrade it in oxygen-rich settings, it persists in natural waters and sediments. PVC is a synthetic polymer that is not biodegradable due to its strong carbon–chlorine bonds and dense structure, which make it resistant to microbial attack. PFOA and PFOS, both part of the PFAS family, are

among the most environmentally persistent chemicals known. Their carbon–fluorine bonds are nearly unbreakable, making them extremely resistant to natural degradation. Based on current research, the biodegradability ranking from highest to lowest is DEP, BPA, PVC , and lastly PFOA and PFOS. As shown in Figure 3, all models classified BPA, PFOA, and PFOS as non-biodegradable, highlighting their persistence and toxicity in accordance with current research. In contrast, DEHP was classified as biodegradable by all models, also supported by current research. However, PVC was classified as biodegradable by all models except KNN. This prediction is inaccurate when compared to current research so the classification of PVC was disregarded. This led to a focus on BPA, PFOA, and PFOS, which were consistently identified as non-biodegradable, to identify potential safer, biodegradable substitutes.

To identify safer alternatives, cosine similarity was used to compare the structures of BPA, PFOA, and PFOS against a dataset of 2,500 naturally occurring chemicals. The top 50 cosine similarity values for each of these three chemical carcinogens were identified. The results of this analysis, shown in Figure 4, revealed a significant difference in the structural similarity between BPA and the naturally occurring chemicals compared to PFOA and PFOS. The average cosine similarity of BPA to the 50 most similar chemicals from the dataset was around 75%, while PFOA and PFOS had an average similarity of only 17%. This low similarity arises because PFOA and PFOS belong to the broader class of synthetic chemicals known as PFAS. In PFAS, fluorine atoms replace hydrogen atoms in traditional hydrocarbon structures, creating a perfluorinated tail. This unique structure gives PFAS distinctive properties such as heat and water resistance. Fluorine–carbon bonds are extremely rare in nature, explaining the low cosine similarity value. Because the natural chemicals were too structurally different from PFOA and PFOS to be considered viable substitutes, the focus of the study was narrowed to finding

potential substitutes specifically for BPA.

Using cosine similarity, structurally similar compounds to BPA were identified and assessed for biodegradability. The results revealed promising alternatives, including 3-Propionyl Chromone, N-Acetyl-D-glucosamine, and Butyl Paraben, which were found to be biodegradable and could serve as potential substitutes for BPA in plastics.

Conclusion

As plastic pollution and chemical exposure continue to pose health and environmental risks, we must develop more efficient solutions to reduce dependence on slow, traditional biodegradation testing. This project created high-performing machine learning models trained with RDKit-generated molecular descriptors. The RDKit-trained models outperformed PCA-based models in predicting the biodegradability of carcinogenic chemicals. Among the tested algorithms, neural network and random forest models delivered the best overall performance and met the design criteria for accurate and reliable classification. The neural network achieved the highest accuracy, precision, and recall scores, along with the lowest false positive rate, while the random forest model recorded the best AUC score.

In addition to classification, this project also aimed to suggest safer alternatives to toxic additives. The analysis identified BPA as a viable candidate for substitution based on cosine similarity with naturally occurring chemicals. The project identified biodegradable alternatives for BPA, including 3-Propionyl Chromone, N-Acetyl-D-glucosamine, and Butyl Paraben. These results demonstrate that machine learning can potentially reduce the time and resources needed to evaluate chemical safety. By applying cosine similarity analysis, we found structurally similar and more sustainable alternatives, offering a data-driven path forward in plastic production.

Overall, this research supports ongoing efforts to mitigate plastic pollution and reduce health risks associated with persistent carcinogenic chemicals.

References

1. Andrady, A. L. (2011). “Microplastics in the marine environment,” *Marine Pollution Bulletin,* 62(8), 1596–1605. https://doi.org/10.1016/j.marpolbul.2011.05.030
2. Cho, J., Kim, C., Lee, Y., Sohn, J. (2023). “Health Effects of Microplastic Exposures: Current Issues and Perspectives in South Korea,” *Yonsei medical journal*, *64*(5), 301–308. https://doi.org/10.3349/ymj.2023.0048
3. González, L. (2024). “I’m a Microplastics Researcher. Here’s How To Limit Their Dangers,” *University of California, San Francisco*. https://www.ucsf.edu/news/2024/02/427161/how-to-limit-microplastics-dangers
4. Hopewell, J., Dvorak, R., Kosior, E. (2009). “Plastics recycling: Challenges and opportunities,” *Philosophical Transactions of the Royal Society B: Biological Sciences,* 364(1526), 2115–2126. https://doi.org/10.1098/rstb.2008.0311
5. Khedkar, S. (2024).”'Very concerning': Microplastics can accumulate in cancer cells and may help them spread, study hints,” *Live Science*. https://www.livescience.com/health/cancer/very-concerning-microplastics-can-accumulate-in-cancer-cells-and-may-help-them-spread-study-hints
6. Li, T., Wu, Y., Yang, L. (2012). “Mechanisms of removing pollutants from aqueous solutions by microorganisms and their aggregates: A review,” *Science Direct.* https://www.sciencedirect.com/science/article/abs/pii/S0960852411018529

7. Organisation for Economic Co-operation and Development (OECD). (2002). *Hazard Assessment of Perfluorooctane Sulfonate (PFOS) and Its Salts*. https://www.oecd.org/chemicalsafety/risk-assessment/2382880.pdf

8. Organisation for Economic Co-operation and Development (OECD). (2002). *SIDS Initial Assessment Report for Bisphenol A*. United Nations Environment Programme (UNEP). https://www.oecd.org/chemicalsafety/risk-assessment/1875526.pdf

9. Organisation for Economic Co-operation and Development (OECD). (2002). *SIDS Initial Assessment Report for Phthalates.* United Nations Environment Programme (UNEP). https://hpvchemicals.oecd.org/UI/handler.axd?id=6ea10452-9a10-4d7e-9121-c55d5d574f

10. Osborne, M. (2022). "Microplastics Detected in Human Blood in New Study," *Smithsonian Magazine*. https://www.smithsonianmag.com/smart-news/microplastics-detected-in-human-blood-180979826/

11. PubChem. (n.d.). *Diethyl Phthalate (DEP) Compound Summary.* National Center for Biotechnology Information. https://pubchem.ncbi.nlm.nih.gov/compound/Diethyl-phthalate

12. Staples, C. A., Dorn, P. B., Klecka, G. M., O'Block, S. T., Harris, L. R. (1998). "A review of the environmental fate, effects, and exposures of bisphenol A," *Chemosphere,* 36(10), 2149–2173. https://doi.org/10.1016/S0045-6535(97)10133-3

13. Staples, C. A., Peterson, D. R., Parkerton, T. F., Adams, W. J. (1997). "The environmental fate of phthalate esters: A literature review," *Chemosphere,* 35(4), 667–749.

https://doi.org/10.1016/S0045-6535(97)00195-1

14. Wang, Z., DeWitt, J. C., Higgins, C. P., Cousins, I. T. (2017). “A never-ending story of per- and polyfluoroalkyl substances (PFASs)?” *Environmental Science & Technology,* 51(5), 2508–2518. https://doi.org/10.1021/acs.est.6b04806

15. Zambrano, M. C., Pawlak, J. J., and Venditti, R. A. (2020). "Effects of chemical and morphological structure on biodegradability of fibers, fabrics, and other polymeric materials," *BioRes*. 15(4), 9786-9833. https://bioresources.cnr.ncsu.edu/resources/effects-of-chemical-and-morphological-structure-on-biodegradability-of-fibers-fabrics-and-other-polymeric-materials/#:~:text=Generally%2C%20the%20biodegradability%20of%20a,(easier%20depolymerization)%20and%20crystallinity