

High Data Quality Enhances Microplastic Toxicity Prediction

Ana L. Antonio Vital¹, Scott Coffin², Andrea Bonisoli-Alquati³, Maaïke Vercauteren⁴,
Luan de Souza Leite⁵, Maximilian Pichler⁶, Magdalena M. Mair^{1,7*}

¹Statistical Ecotoxicology, University of Bayreuth, Bayreuth, Germany

²California State Water Resources Control Board, Sacramento, USA

³California State Polytechnic University, Pomona, USA

⁴Blue Growth Research Lab, Ghent University, Ghent, Belgium

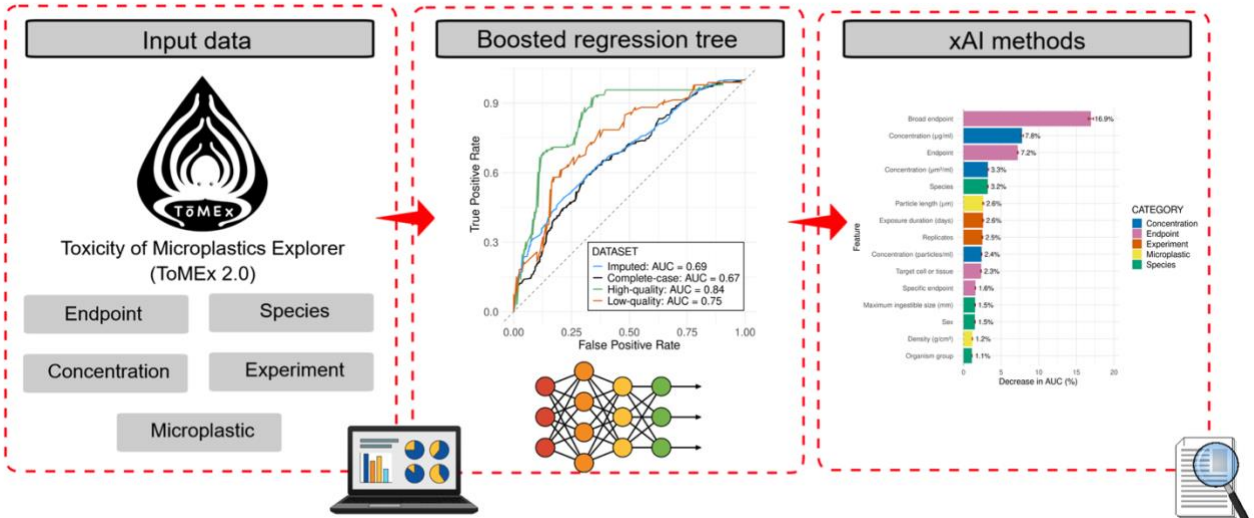
⁵State University of Campinas (UNICAMP), Campinas, Brazil

⁶Group for Theoretical Ecology, University of Regensburg, Regensburg, Germany

⁷Bayreuth Center for Ecology and Environmental Research (BayCEER), Bayreuth,
Germany

*Corresponding author: magdalena.mair@uni-bayreuth.de

17 Graphical Abstract



18

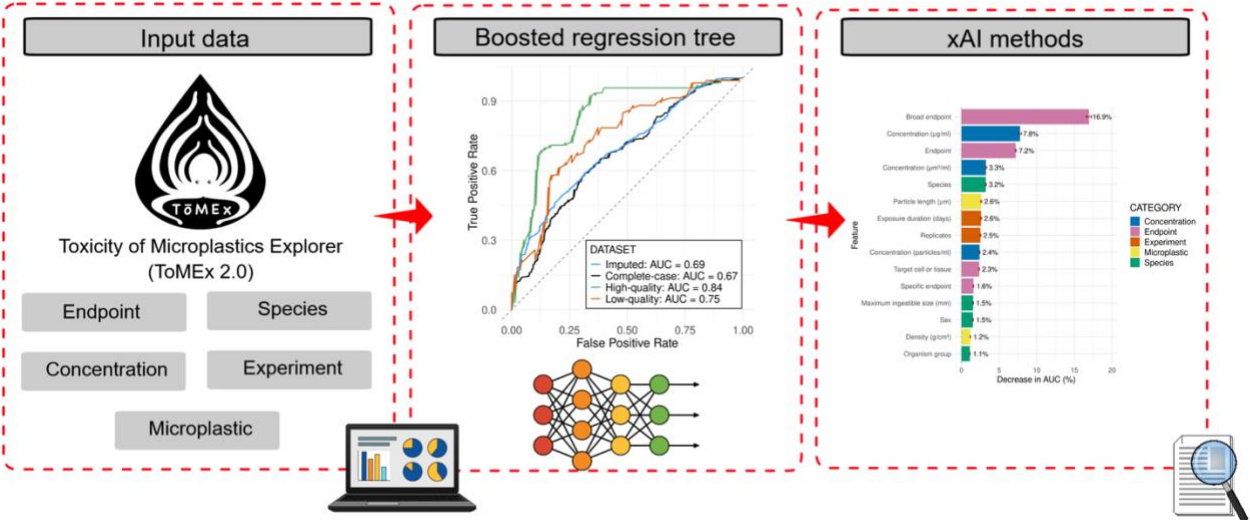
19

Abstract

Unlike chemicals, microplastic (MP) particles lack standardized identifiers, limiting the applicability of traditional predictive ecotoxicology methods such as quantitative structure-activity relationship (QSAR) models. By leveraging an extensive database of MP effects, this study aimed to predict MP toxicity using MP properties, MP concentration, organismal traits, endpoints, and experimental design as features, and to evaluate how data pre-processing, dataset size, and data quality influence model performance. Specifically, we applied the Boosted Regression Tree (BRT) machine learning algorithm to four datasets derived from the Toxicity of Microplastics Explorer database (ToMEx 2.0): (i) imputed missing values, (ii) complete-case (missing values removed), (iii) high-quality data, and (iv) low-quality data. The high-quality dataset yielded the best final predictions with highest area under the curve (AUC) values for both random cross-validation (AUC = 0.93) and blocked cross-validation by particle identifier schemes (AUC = 0.87). Explainable artificial intelligence (xAI) analyses showed that predictive performance was primarily determined by endpoints and concentration, with contributions by MP properties despite limited reporting. Our findings demonstrate the feasibility of predicting MP toxicity and identifying key drivers for its predictive performance. They further indicate that high-quality data improve predictive performance, reducing the need for extensive data mining efforts and computational costs. Standardized experimental design, detailed characterization of MP properties, and rigorous reporting practices will enhance toxicity predictions, support more reliable risk assessment frameworks, and inform the design of safer materials in the future.

Keywords: ecotoxicology, explainable artificial intelligence, predictive modeling, microplastic properties, risk assessment

44 **Graphical Abstract**



45

46

1. Introduction

Microplastics (MPs; < 5 mm) are a complex and multidimensional group of pollutants (Rochman *et al.*, 2019; Monclús *et al.*, 2025). Each MP particle has its own set of physicochemical properties, or “traits” related to shape, size, polymer type, surface charge, eco-corona, and plastic-associated chemicals, including additives (Hartmann *et al.*, 2019; Lambert *et al.*, 2017). Moreover, MP properties change under degradation or weathering processes due to exposure to the natural environment (*e.g.*, temperature, UV light, and wind; Menzel *et al.*, 2022; Ramsperger *et al.*, 2020). Studies have shown that properties such as size (Abarghouei *et al.*, 2021), shape (Schwarzer *et al.*, 2022), surface characteristics (Wieland *et al.*, 2024), and polymer type (Hodkovicova *et al.*, 2022) significantly contribute to toxicity by influencing bioavailability, biological interactions, and the severity of adverse effects upon exposure (Wieland *et al.*, 2022). To understand the effects of MP exposure, it is essential to consider their diverse physicochemical properties and interactions

Additionally, experiments on MP toxicity are often time-consuming and extensively reliant on animal testing, particularly in risk and hazard assessment contexts. A comprehensive hazard assessment of multiple MPs and their combined properties across different species and endpoints would require testing an impractically large number of MP property combinations across all relevant species and endpoints, making such an approach inefficient and resource-prohibitive (Koelmans *et al.*, 2023). Similarly, it is increasingly recognized that testing every plastic-associated chemical in a laboratory is infeasible and ethically undesirable given the large number of plastic chemicals, estimated at over 16,000 (Monclús *et al.*, 2025).

This challenge has driven the implementation of New Approach Methodologies (NAMs) for chemical hazard assessments, which, in a broader context, include *in-chemico*,

72 *in-silico*, and *in-vitro* approaches (European Chemicals Agency, 2016). NAMs aim to
73 reduce animal testing and enable more cost- and time-efficient chemical assessments.
74 Chemicals can be linked to unique identifiers (*e.g.*, SMILES and CAS numbers) through
75 their well-defined molecular structures and stable properties, which enables the use of *in-*
76 *silico* NAMs by ensuring the integrity and reproducibility of chemical information across
77 databases and publications (Müller, 2025). Within the NAM frameworks, particularly
78 quantitative structure-activity relationship (QSAR) modeling has demonstrated robust
79 predictive capabilities for chemicals by establishing mathematical correlations between
80 chemical structures and toxic outcomes (Tropsha, 2010).

81 Machine learning (ML) methods have emerged as a promising alternative to
82 traditional QSAR modeling approaches. More generally, ML techniques employ a variety
83 of algorithms to identify patterns in data and generate predictions (Pichler & Hartig,
84 2023b). First approaches have been published that apply ML-based modeling to predict
85 MP toxicity by linking MP properties to observed effects (Wang *et al.*, 2026; Zhang *et al.*,
86 2025). The integration of plastic material properties into these models can support the
87 development of more environmentally safer plastic materials, thereby reducing the need
88 for animal testing in material development. Furthermore, by revealing the current state of
89 evidence, a predictive approach to MP toxicity can identify data gaps, highlighting areas
90 that require further investigation and support risk assessment frameworks. However, a
91 limitation of deploying ML algorithms is their lack of interpretability (Rudin, 2009).
92 Unlike statistical models, ML algorithms generally do not produce coefficients that clearly
93 indicate relationships between predictor traits and toxicity. Fortunately, explainable
94 artificial intelligence (xAI) tools help us understand how ML algorithms make their
95 predictions (Renftle *et al.*, 2024), potentially allowing the identification of MP properties
96 or experimental conditions that drive toxicity outcomes.

While the idea of using ML to predict MP toxicity from MP properties is appealing, thus far only a few studies have used ML approaches to predict the toxicity of MP (*e.g.*, Choi *et al.*, 2023; Li *et al.*, 2025; Wang *et al.*, 2026; Zhang *et al.*, 2025; Zhen *et al.*, 2023). Most studies on MP research have used ML methods for identification (Gong *et al.*, 2023; Meyers *et al.*, 2024), detection (Devipriya *et al.*, 2025; Jin *et al.*, 2024; Khanam *et al.*, 2025), or MP quantification tasks (Tran *et al.*, 2023). In studies predicting MP toxicity, the focus is often on human health, with predictions relying on limited datasets on MP effects on tissues or cell cultures (*e.g.*, Choi *et al.*, 2023; Zhen *et al.*, 2023). Recently, one study (Zhang *et al.*, 2025) applied ensemble ML to explore factors influencing the combined toxicity of plastic-pollutant mixture across different species, using data from 41 studies (351 data points), including nine input features describing the physicochemical properties of MPs (*e.g.*, polymer type, shape, size) and experimental conditions (*e.g.*, exposure duration and concentration). However, the domain of applicability remained narrow, with only three polymer types (*i.e.*, polystyrene, polyethylene, and others) and one endpoint (*i.e.*, survival rate) included, providing a valuable but limited basis for generalizing the results (Zhang *et al.*, 2025). Another recent publication (Wang *et al.*, 2026) used random forests as the top-performing ML algorithm to predict NOEC (no observed effect concentration) values from MP toxicity data across aquatic species. Still, the study included only five polymer types (*i.e.*, polystyrene, polyethylene, polypropylene, polyethylene terephthalate, and polyvinyl chloride) and seven predictor variables: polymer type, shape, length, density, species group, endpoint, and environment type.

To increase the environmental relevance of these predictive modeling approaches for MP toxicity, larger and more ecologically diverse datasets are needed, including information from multiple species, endpoints, and a broad range of MP particles with variable physicochemical properties (Yu *et al.*, 2022). To date, the largest publicly

available database covering this diversity is the Toxicity of Microplastics Explorer (ToMEx 2.0) database (Thornton Hampton *et al.*, 2022; Thornton Hampton *et al.*, 2025). As part of its more recent extension, nearly 60 scientists worldwide mined and validated 6,927 data points from 124 publications, representing approximately two years of research on the effects of MPs on aquatic organisms. This effort expanded the ToMEx database to almost 13,000 data points derived from 286 studies.

With the growing number of publications on MP research, there have been increasing calls to standardize key aspects of MP toxicity studies, which is also pivotal for the development of robust QSAR models for MP toxicity. In response, quality criteria were proposed (de Ruijter *et al.*, 2020; Mehinto *et al.*, 2022) to better align MP research with established standards in other ecotoxicological fields (*e.g.*, chemical testing). To identify the relevant datasets required for the development of preliminary risk assessments and threshold values, the ToMEx 2.0 database incorporates data scores based on de Ruijter *et al.* (2020) and Mehinto *et al.* (2022). Within ToMEx 2.0, each experiment is rated according to the minimum technical and risk-assessment quality criteria to determine whether it is sufficiently reliable and informative for hazard assessment (Kennedy *et al.*, 2025). However, to date, the importance and effectiveness of data quality in predictive models of MP toxicity remain poorly understood.

In this study, we integrate information from 28 polymer types and 44 features related to MP properties, species, endpoint, and experimental design to predicted MP toxicity. We utilized xAI approaches to identify the most important features for toxicity prediction and thus identify key drivers of MP toxicity outcomes. For this task, and to understand how dataset size, data quality, and data pre-processing approaches influence the predictions, we used four different datasets derived from the ToMEx 2.0 database: (i) an imputed dataset (*i.e.*, where we imputed missing values); (ii) the complete-case dataset

(i.e., with missing values removed); (iii) a high-quality dataset; and (iv) a low-quality dataset. Each dataset was evaluated using two different cross-validation schemes: (i) stratified random cross-validation and (ii) blocked cross-validation by particle identifier (i.e., particle ID). We hypothesized that predictors such as MP size and concentration, as well as species- and endpoint-related features, significantly contribute to toxicity prediction, and that the high-quality dataset would likely yield the best predictive performance.

2. Material and methods

Data source

As the basis for our predictive models, we used the ToMEx 2.0, an open-source database that aggregates published data on the effects of MPs on aquatic organisms, along with their corresponding data quality scores (Thornton Hampton *et al.*, 2025; Thornton Hampton *et al.*, 2022). This database comprises nearly 13,000 data points extracted from 286 published papers. The ToMEx 2.0 database is the result of an international crowdsourcing effort that extended ToMEx 1.0 by mining and validating literature on MP toxicity to aquatic organisms published between January 1, 2021, and January 11, 2023 (Thornton Hampton *et al.*, 2025). While standard toxicology metrics (e.g., NOEC or LC50 values) offer a more practical understanding of MP toxicity, and effect size estimations a more quantitative one, effect reporting in ToMEx 2.0 is limited to a binary response variable as either observed (yes = 1) or not observed (no = 0), based on whether the treatment containing MP particles statistically differed from the particle-free control based on dependent on reported p-values provided by the authors.

Data preparation

To increase taxonomic consistency in the dataset, we removed data points on plants from all four datasets (see File S2). In ToMEx 2.0, several data points frequently correspond to nominally identical particles. Since we also aim to predict particles for which no prior information is available (*i.e.*, particles not included in the training phase), we created particle identifiers (particle IDs) based on unique combinations of DOI, polymer type, particle shape, and particle length (μm), which were used to guide the data split for the blocked cross-validation scheme.

Where possible, missing values were first manually imputed by selecting the most reasonable estimates: Missing temperature values were filled up using the median temperature for each species, assuming that temperatures in the experimental protocols would be similar within each species. If no data were available for the same species, the remaining missing temperature values were filled in using the median temperature of the higher organism group (*i.e.*, higher taxonomic level, *e.g.*, crustacea, fish, insects). Particles with lengths $\geq 1000 \mu\text{m}$ were flagged for manual review before imputation. The corresponding publications were reviewed to confirm reporting accuracy; corrections were made if necessary, and corrected values were incorporated into the ToMEx 2.0 database. If the particle width (μm) was missing, but particle length (μm) was reported and the MP shape was spherical, the width was assumed to be equal to the reported length. For particle-only studies (*i.e.*, no co-exposure), missing values for the chemical dose (mg/L) were set to zero. When sample sizes were reported as concentrations within the test vessels, or were missing because they were uncountable (*e.g.*, for organism groups such as algae, bacteria, cyanobacteria, and dinoflagellates), sample size was set to $n = 1$ to reflect the single test vessel from which the measurements were derived. For missing polymer density values, we first added the median density (g/cm^3) for each polymer type with available values. The remaining missing values were then filled with literature-based densities for specific

polymers. Factor levels were collapsed and harmonized where necessary for the following factorial features: life stage, sex, exposure route, particle mix, exposure media, solvent, detergent, chemicals added, effect, direction, target cell or tissue, polymer, particle source, sodium azide present, particle cleaning, solvent rinse, particle behavior, organism fed, and weathered or biofueled. Variables not relevant for prediction (*e.g.*, DOI, year, authors) or related to quality scores (*e.g.*, particle size score, polymer type score) were dropped. Relevant features with at least 50% observed values were retained as predictors (Tang & Ishwaran 2017). Consequently, no more than half of the values for any given feature were imputed, thereby minimizing the introduction of noise. For a complete list of features used and dropped, along with their corresponding missing value percentages, please see File S2

To evaluate how data pre-processing, dataset size, and data quality influence the predictive performance of the models, we derived four datasets using different pre-processing approaches from the ToMEx 2.0 database. (*i*) The first dataset (imputed dataset; 11,404 data points, 261 studies) included all data points from the ToMEx 2.0 aquatic database after our data preparation described above. In the imputed dataset, the remaining missing values were imputed separately for two feature groups using the *missRanger* package (Mayer, 2024), which implements random-forest-based imputation by estimating the most likely values for each missing observation. The first feature group consisted uniquely of features related to experimental characteristics (17 features, *e.g.*, exposure duration, sample size, endpoint). This feature group was used because we expected that experiments that resemble each other in their reported features would also resemble each other in non-reported features. The second feature group consisted of stressor-related features (20 features, *e.g.*, particle length, particle shape, solvent rinse). The multiple imputations were performed on the training subset (see section ‘Data split’ below). To avoid data leakage (*i.e.*, the leakage of information from the training to the validation and

test datasets), missing values in the validation and test subsets were imputed using the random-forest-based model previously trained on the training subset. (ii) For the second dataset (complete-case dataset), we removed all samples that contained at least one missing value in any of the features. This resulted in a much smaller dataset (4,514 data points; 128 studies). (iii) The complete-case dataset was then further filtered according to the quality criteria established by Mehinto *et al.* (2022) and de Ruijter *et al.* (2020). Only data points from experiments meeting the minimum technical and risk assessment quality criteria requirements were kept (high-quality dataset; 2,038 data points; 54 studies). For a comprehensive understanding of the quality criteria thresholds, refer to Thornton Hampton *et al.* (2022, Table S2). (iv) Finally, for comparison, we created a fourth dataset (low-quality dataset; 1,816 data points, 84 studies), including only samples that failed to meet the minimum quality criteria, for either technical or risk assessment. For this last dataset, we expected lower predictive performance than for the other datasets.

Data split

Random cross-validation assumes that all data points are independent, whereas blocked cross-validation assumes that data points are grouped (*e.g.*, by particle) and that the entire group should be assigned to the same cross-validation fold to accurately reflect predictions for unseen groups. Consequently, blocked cross-validation provides a more realistic scenario when the goal is to generalize to new particles. Considering that, we tested and compared both cross-validation schemes. For the random cross-validation, each of the four datasets was split into three stratified subsets (training, validation, and test) while preserving the distribution of the binary response variable (*i.e.*, toxicity: yes = 1, no = 0) in each split (see Table S1). Each dataset (imputed, complete-case, high-quality, and low-quality) was partitioned into training, validation, and test subsets, keeping an approximate

ratio of 60% training, 20% validation, and 20%. For the blocked cross-validation, in addition to the rough stratification based on the binary response variable, the split was performed at the particle ID level to ensure that all measurements from the same particle were assigned to only one subset (*i.e.*, either training, validation, or test). This approach avoids the presence of identical particles in more than one data subset. For the blocked cross-validation, each dataset (*i.e.*, imputed, complete-case, high-quality, and low-quality) additionally maintained an approximate 60:20:20 split, respectively for training, validation, and testing, according to the total number of particle IDs (see Table S2).

Modeling and evaluation

To predict the response variable (*i.e.*, MP toxicity), we employed the decision-tree-based Boosted Regression Tree (BRT) method. This efficient ML algorithm is suitable for both regression and classification tasks and improves predictive performance by sequentially fitting several models in an iterative process (Elith *et al.*, 2008). During the development phase, we conducted exploratory modelling using Random Forest, Neural Network, and BRT algorithms. BRTs were selected not only for their superior performance during validation, but also for their recognized reliability in identifying strong predictors and detecting patterns that might be causal (Pichler & Hartig, 2023a, 2023b), which directly aligns with our aim of identifying properties associated with MP toxicity. Additionally, it is widely accepted in ML research that classical ML algorithms, such as BRT (in particular, extreme boosting), achieve higher accuracy on tabular datasets compared to deep learning approaches. As the performance of these algorithms depends strongly on their hyperparameters (*e.g.*, regularization for complexity control), we performed hyperparameter tuning using a random grid of 100 parameter set combinations, with first ranges defined based on prior knowledge (see the list of parameters in Table S3).

Model performance was evaluated using a 5-fold cross-validation on the training subset. The training data were randomly divided into five approximately equal folds. In each iteration, one fold served as the validation set, while the remaining four folds were used to train the model. This process was repeated five times so that each fold served as the validation set once. For the blocked cross-validation scheme, the folds were defined by particle ID. Particle IDs were partitioned into five groups, ensuring that all observations belonging to a given particle remained within the same fold. After data splitting into folds, the particle IDs were removed and not used as a predictive feature. The performance metrics were averaged across all the five iterations under each cross-validation strategy. This process was repeated for each of the 100 hyperparameter combinations. The best-performing hyperparameter set was used to fit a BRT model to the training subset. This model was then used to predict effects on the validation subset. If performance on the validation subset continued to improve after tuning, we adjusted the hyperparameter range and repeated the entire procedure. After optimizing model tuning (*i.e.*, predictive performance on the validation subset was satisfactory), the final model was fitted using the best hyperparameters on the combined training and validation subsets. This final model was subsequently used to predict MP toxicity in the independent test subset and to assess its final predictive performance.

To evaluate the model performance (including hyperparameter tuning), we used Receiver Operating Characteristic (ROC) curves and their respective AUC (area under the ROC curve) values as measures of predictive performance. AUC values measure the ability of a model to distinguish between classes, in our case, MP toxicity observed (yes = 1) or not observed (no = 0) by ranking the continuous probabilities. Similarly, we calculated the Precision-Recall Area Under the Curve (PR-AUC). By plotting precision against recall, the PR-AUC focuses on the positive instances, being especially meaningful for binary

classification when the positive class is rarer. Both ROC-AUC and PR-AUC range from 0 to 1, where 1 indicates a perfect predictive performance; 0.5 corresponds to random guessing, and 0 suggests completely incorrect predictions. Furthermore, to isolate the potential influence of dataset size, an additional analysis was conducted in which the imputed, complete-case, and high-quality datasets were randomly downsampled to match the size of the smallest dataset (*i.e.*, the low-quality dataset, containing 1,816 data points) and restricted to only common variables across all datasets after one-hot-encoding (*i.e.*, after converting categorical data into numerical vectors). For this analysis, the best hyperparameters identified during the blocked cross-validation tuning process were utilized. The data as portioned into training (80%; $n = 1,453$) and testing (20%; $n = 363$) sets across 1,000 random iterations. For each iteration, the AUC was calculated, and the average AUC across all 1,000 runs was computed. Additionally, we calculated the Matthew's Correlation Coefficient (MCC) on the final model (*i.e.*, for the best threshold) to evaluate whether the AUC value is inflated by the imbalance on the predictor variable (effect yes = ~35%; effect no = 65%). Unlike other metrics (*e.g.*, accuracy), which can be misleading when one class is more frequent than the other, MCC only scores high when both classes perform well. The MCC values range from -1 (total incorrect classification) to 1 (perfect classification), with 0 indicating performance no better than random guessing.

Feature importance

To identify the features driving MP toxicity prediction, we applied xAI methods to the high-quality dataset using the best-performing model. Permutation importance for features was derived from the blocked cross-validation model to understand how input features contribute to model predictions (Fisher *et al.*, 2019). Permutation importance was assessed by comparing the AUC of the baseline model (*i.e.*, predictions on the full dataset)

with the AUC obtained when a feature's values were randomly shuffled. The shuffling process was repeated 20 times for each feature, and the average percent reduction in AUC relative to the baseline AUC was calculated to assess the drop in model performance when feature values were randomized.

To infer potential feature-feature (trait-trait) interactions among numerical features, we used the xAI Average Conditional Effect (ACE) method to estimate pairwise feature interactions from the blocked cross-validation model. Conditional effects quantify how model predictions change when one predictor varies while all the others remain at their observed values. Unlike classical marginal effects, which are based on unit changes, conditional effects are computed using small changes in the predictor (*i.e.*, finite difference). These local effects are then averaged across observations to obtain an average conditional effect for each predictor (*i.e.*, the global effects). Two-way conditional effects are computed by jointly perturbing two predictors by small amounts.

To determine whether MP properties are important for predicting MP toxicity or whether toxic outcomes are driven primarily by concentration or experimental design, we investigate the contribution of each feature category (*i.e.*, concentration, endpoint, experiment, MP, and species) using a feature-group ablation analysis. We used the best model from both random and blocked cross-validation schemes to train models under 10-fold cross-validation with different combinations of the five feature categories. Under the blocked cross-validation, each particle ID was assigned to only one of the 10 folds. For each model, the mean and standard deviation (SD) of the AUC values across folds were recorded. All data modeling and visualizations were performed in R (version 4.4.1; R Core Team, 2024).

3. Results

Overall, our models demonstrated that predicting MP toxicity using the selected features is feasible, despite the high heterogeneity in the ToMEx 2.0 data. Predictions on the test subset using the random cross-validation scheme (Figure 1A) achieved the best predictive performance with the high-quality dataset (AUC = 0.93), the only one to reach an AUC value above 0.9, followed by the imputed (AUC = 0.89), complete-case (AUC = 0.87), and low-quality dataset (AUC = 0.83). The blocked cross-validation (Figure 1B) also performed best with the high-quality dataset (AUC = 0.84), followed by the low-quality dataset (AUC = 0.75), the imputed dataset (AUC = 0.69), and the complete-case dataset (AUC = 0.67). Additional AUC values from the training and validation phases are provided in Tables S4 and S5. Under random cross-validation (Figure 1C), PR-AUC performance followed the order: high-quality (PR-AUC = 0.91), imputed and complete-case (PR-AUC = 0.82), and low-quality (PR-AUC = 0.79). Under blocked cross-validation (Figure 1D), PR-AUC was highest for the high-quality dataset (PR-AUC = 0.74), followed by complete-case (PR-AUC = 0.64), low-quality (PR-AUC = 0.63), and imputed (PR-AUC = 0.52). The AUC values of the models with datasets reduced to the same size (Figure S1) followed a similar trend. Additional MCC values are reported in Table S6.

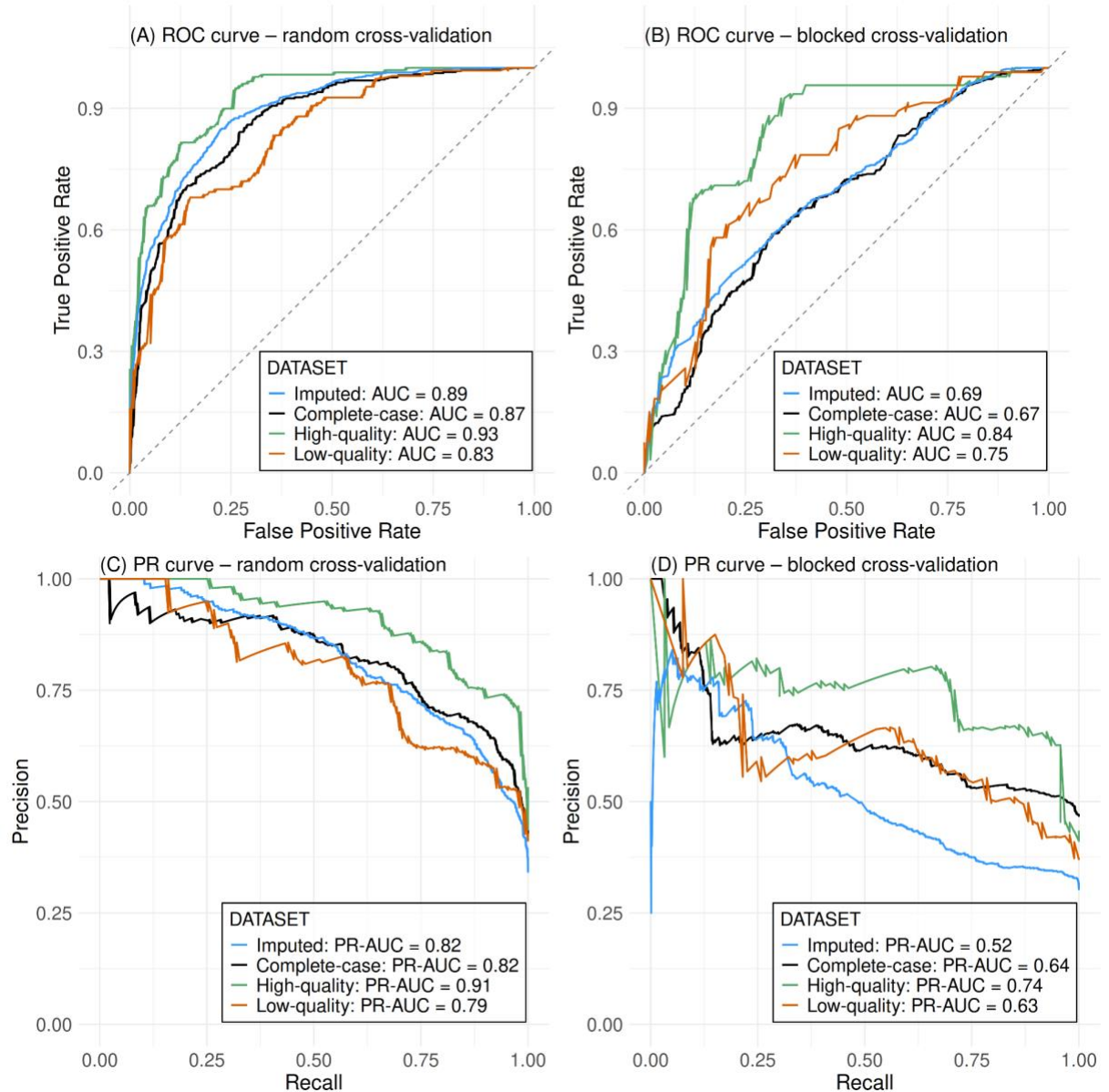


Figure 1. Comparison of the performance of models trained on imputed, complete-case, high-quality, and low-quality datasets. (A) and (B) show ROC (receiver operating characteristic) curves and AUC (area under the curve) values derived from predictions on the independent test subset for both cross-validation schemes. (A): ROC-curve under random cross-validation. (B): ROC-curve under blocked cross-validation. (C) and (D) show area under PR-AUC (precision-recall area under the curve) curves derived from predictions on the independent test subset for both cross-validation schemes (C): PR-curve under random cross-validation. (D): PR-curve under blocked cross-validation. The diagonal line (dashed grey) represents a random classification for reference.

The permutation importance analysis for the top 15 features (Figure 2) identified the broad endpoint (e.g., fitness, metabolism) as the most significant predictor, resulting in

a 16.9% decrease in AUC. Still within the endpoint category, the endpoint (*e.g.*, population size, catalase activity) with an AUC decrease of 7.2%, the target cell or tissue (2.3%) and the specific endpoint (*e.g.*, reproduction, oxidative stress; 1.6%) additionally appeared. Regarding concentration metrics, all the included metrics: concentration in mass ($\mu\text{g/mL}$; 7.8%), concentration in volume ($\mu\text{m}^3/\text{mL}$; 3.3%), and concentration in number of particles (particles/mL; 2.4%) appeared in the top 15. Within the experiment category, contributions of exposure duration in days (2.6%) and replicates (2.5%) were found. Among species-related features, species (3.2%), maximum ingestible size (mm; 1.5%), sex (1.5%), and organism group (1.1%) additionally contributed to the predictive performance. Two microplastic-related features: particle length (μm ; 2.6%) and density (g/cm^3 ; 1.2%) complete the top 15 ranking.

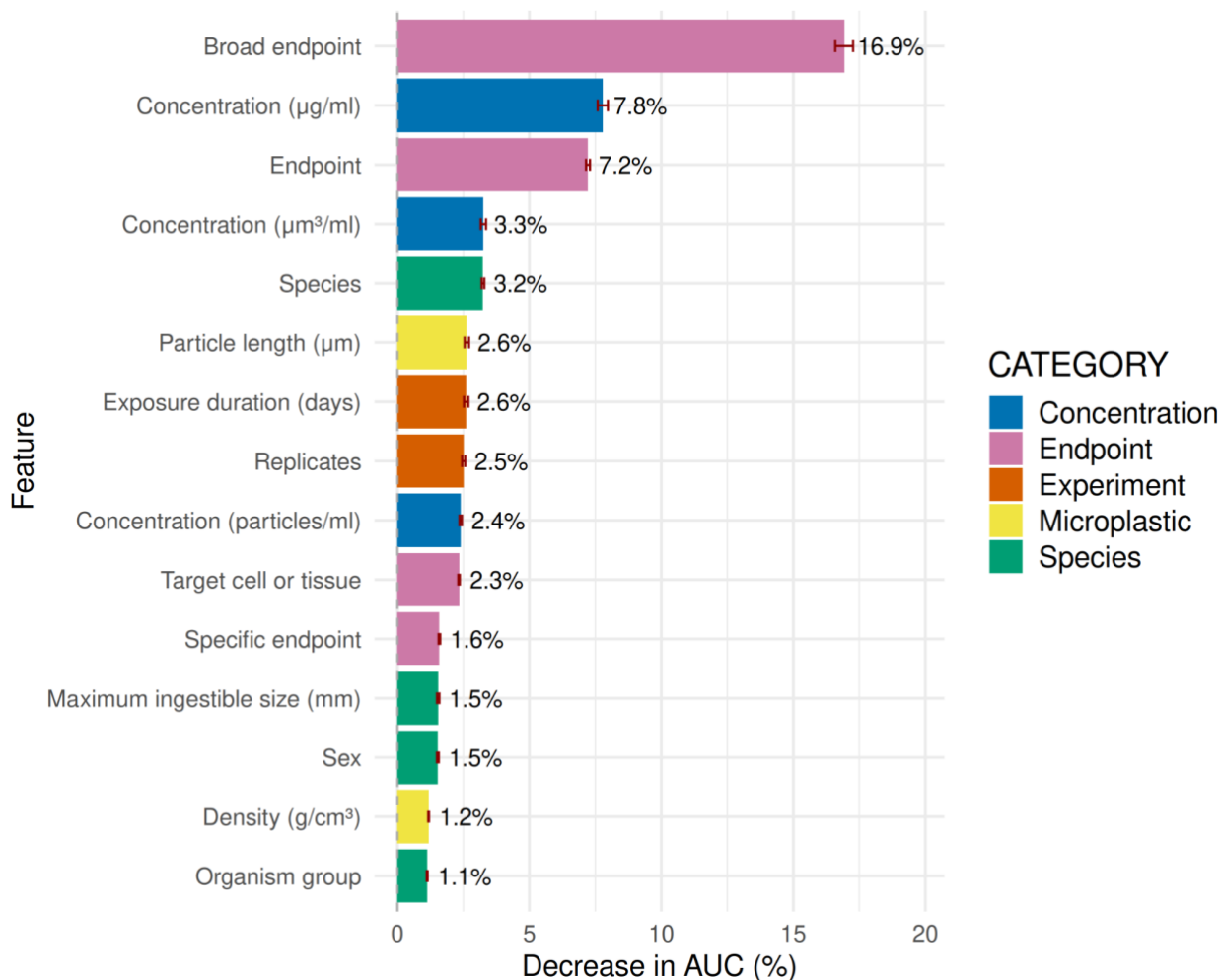


Figure 2. Top 15 decreases in AUC (area under the curve) relative to the baseline model after 20 random shuffles per feature for the final model on the high-quality dataset under the blocked cross-validation scheme. Colors indicate assignments of features to feature categories: concentration (blue), endpoint (pink), experiment (orange), microplastic (yellow), and species (green). Horizontal black bars correspond to the 95% confidence interval (CI).

The Average Conditional Effect (ACE) of pairwise interactions (Figure 3) for numeric features shows how interactions between pairs of features increase (blue) or decrease (red) the probability of observing a toxic outcome. The strongest negative interaction was observed between particle concentration ($\mu\text{m}^3/\text{mL}$) and mass concentration ($\mu\text{g/mL}$). In contrast, the strongest positive interaction was observed between particle length (μm) and concentration in mass ($\mu\text{g/mL}$). The particle length (μm) also showed a strong positive interaction with the estimated maximum ingestible size (mm).

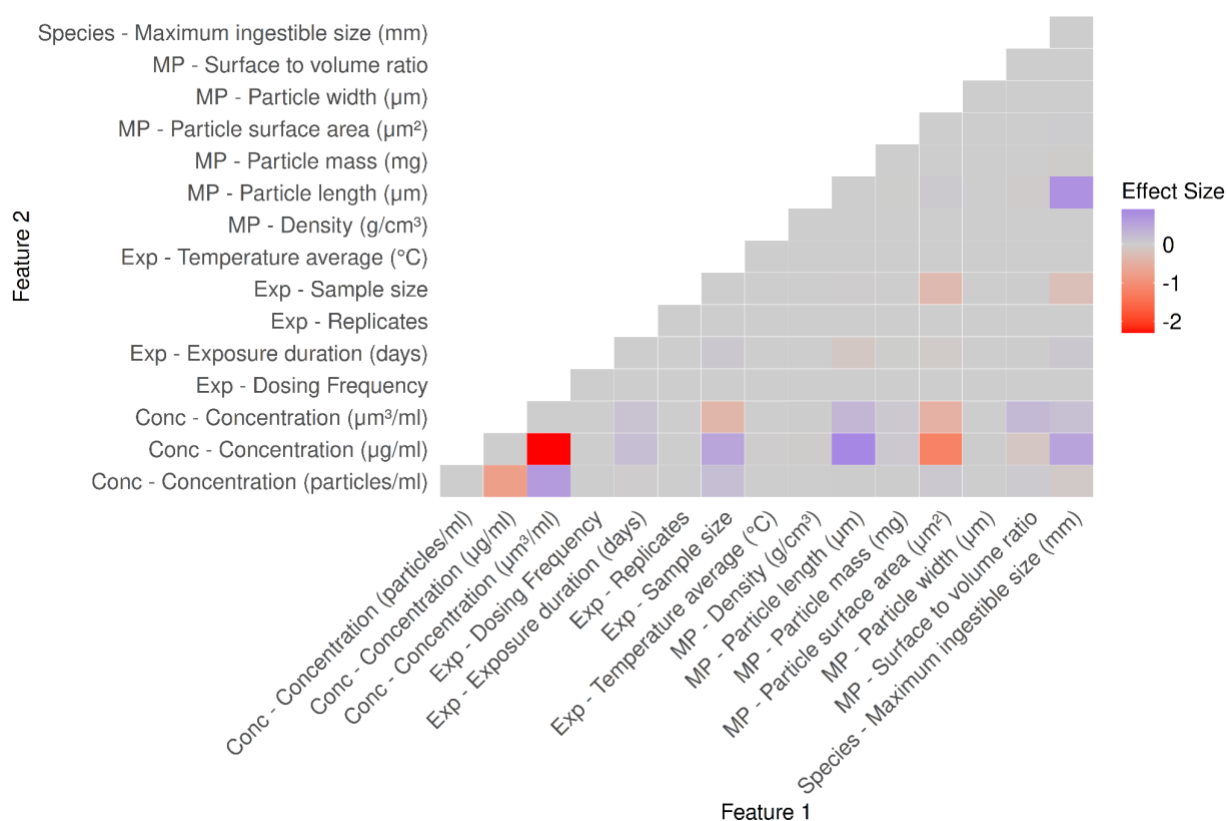


Figure 3. Heat map showing the pairwise interaction strengths of the average conditional effect (ACE) for the high-quality dataset fitted with the hyperparameters from the blocked cross-validation scheme. Interactions in blue increase, and those in red decrease, the probability of observing a toxicity effect. Each feature belongs to one of the following categories: Conc (concentration), Exp (experiment), MP (microplastic properties), and Species (species-related traits).

The feature-group ablation analysis (Figure 4) to assess whether MP properties or other feature categories are more relevant for toxicity prediction shows that, overall, all feature groups (*i.e.*, concentration, endpoint, experiment, MP, and species) contribute to the predictive model's performance. Concentration and endpoint are the most critical groups, particularly when not included as a single group predictor. Under the random cross-validation scheme (Figure 4A; Table S7), the model's best performance included all categories (0.908 ± 0.026), followed by models with four categories, excluding either species (0.906 ± 0.019) or experiment (0.900 ± 0.014). For ablation analysis on single-category models, endpoint presented the highest performance (0.815 ± 0.032), followed by

experiment (0.793 ± 0.031), species (0.772 ± 0.031), concentration (0.759 ± 0.029), and MP properties (0.759 ± 0.042). Under the blocked cross-validation scheme (Figure 4B; Table S8), the SDs are larger and overlap more strongly across different feature combinations. The highest average value includes all the five categories (0.791 ± 0.106); followed by the groups with four categories: concentration, endpoint, experiment, and MP (0.727 ± 0.088), and concentration, endpoint, experiment, and species (0.762 ± 0.092). Among the single-category models, endpoint ranked highest (0.753 ± 0.124), succeeded by experiment (0.693 ± 0.129), species (0.688 ± 0.092), MP (0.614 ± 0.143), and concentration (0.578 ± 0.044).

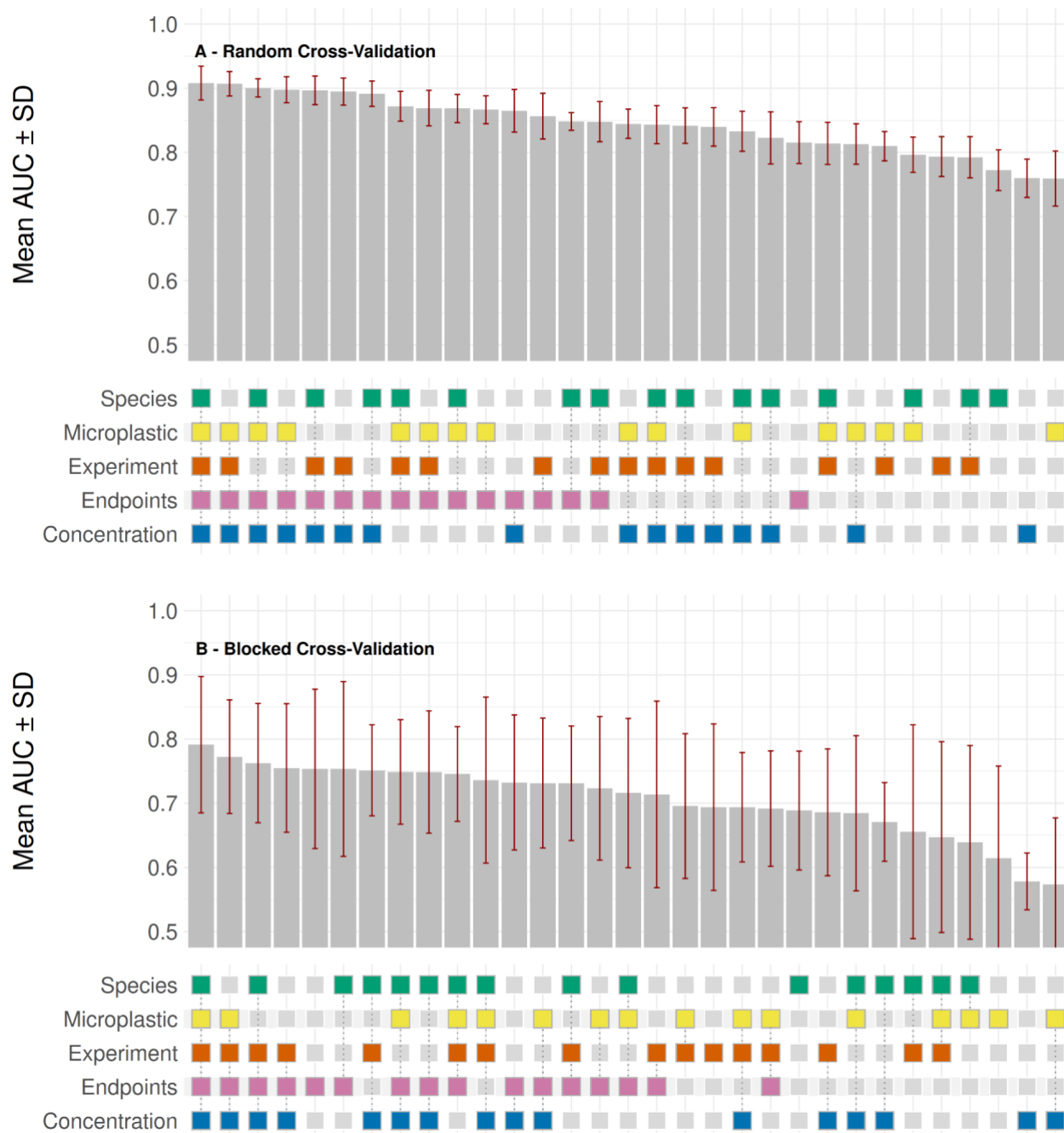


Figure 4. Barplot showing the mean AUC $\pm$ standard deviation for the high-quality dataset across models with different feature categories included and evaluated using 10-fold cross-validation. A: Random cross-validation. B: Blocked cross-validation. The inclusion of distinct feature sets in each model is indicated below its corresponding bar by colored squares: species (green), microplastic (yellow), experiment (orange), endpoints (pink), and concentration (blue).

4. Discussion

Despite the inherent high heterogeneity in the data, our models successfully predicted MP toxicity using the data provided on ToMEx 2.0 database. When comparing

the two cross-validation strategies, the random scheme (AUC = 0.930) outperformed the blocked scheme (AUC = 0.870). This performance decrease under blocked cross-validation was expected, as the model was required to predict the toxicity of new MP particles entirely absent during the training phase. In contrast, random cross-validation allows observations from the same experiment to be split between training and testing sets, which likely lead to data leakage and an overestimation of model performance. Since applying predictive modelling in ecotoxicology requires reliable toxicity estimations for novel substances or materials, the blocked cross-validation scheme provides a more rigorous and realistic evaluation of model generalizability.

Furthermore, our analysis highlights that data quality is more relevant than data volume. The high-quality dataset (~2,000 data points) consistently outperformed the much larger imputed (>11,000 data points) and complete-case (>4,000 data points) datasets. This trend also persisted when all datasets were downsampled to the size of the smallest dataset (*i.e.*, low-quality; ~1,800 data points), confirming that data quality, including factors such as bias and noise (Mohammed *et al.*, 2025) are notably important. In consequence, future efforts in ecotoxicology should prioritize high-quality data, well-characterized experiments and MP particles, and comprehensive and transparent data reporting to maximize predictive performance while minimizing data mining efforts and computational costs (Kennedy *et al.*, 2025).

The positive interaction between particle length and maximum ingestible size suggests that MP toxicity is critically dependent on size-matching between the particle and the organism's physiological capacity for ingestion. This relationship aligns with previous studies demonstrating that particle uptake is highly dependent on the feeding capacity and morphology of the organism (Başaran Kankılıç *et al.*, 2023; Scherer *et al.*, 2017). Furthermore, the interaction between mass concentration and particle length indicates that

larger particles at higher mass might be more toxic, likely through physical mechanisms such as gut damage or intestinal blockage (Kim *et al.*, 2022), leading to reduced feeding efficiency, false satiation and starvation (Lei *et al.*, 2018; Wright *et al.*, 2013).

Permutation importance and feature-group ablation analyses indicated that, although information on MP properties is limited, the included properties, mainly particle length and density, contribute to MP toxicity prediction (Figures 2 and 4). Based on currently availability data from published studies, associations between toxic outcomes and MPs are limited to a few commonly reported MP properties (*i.e.*, shape, density, and dimensions). Plastic-associated chemicals, such as additives and non-intentionally added substances, are often unknown or unreported in the published literature but have been shown to contribute to MP toxicity in several studies (Schränk *et al.*, 2019; Seewoo *et al.*, 2023; Yu *et al.*, 2024). Even nominally identical MP particles can vary substantially in their toxicology (Ramsperger *et al.*, 2022), influenced by frequently unmeasured and unreported properties, such as zeta potential and surface charge (Wieland *et al.*, 2024). Despite these limitations in MP research, a few studies have shown that meaningful toxicity predictions can still be derived from a small set of the commonly reported MP properties (*e.g.*, Wang *et al.*, 2025; Zhang *et al.*, 2025).

More specifically, we found that particle size, measured as length, was the most important MP property for toxicity prediction (Figure 2), which aligns with previous publications (Wang *et al.*, 2025; Zhang *et al.*, 2025). Wang *et al.* (2025) found that MP length was the most important feature in their model. Consistent with the findings of Zhang *et al.* (2025), particle size, measured as diameter, was also identified as the most important MP property for toxicity prediction. In ecotoxicological research, the size of MP has been associated with adverse effects (Cui *et al.*, 2024), with smaller particles generally inducing higher toxicity (Wieland *et al.*, 2022). This relationship was observed across several

endpoints and species (An *et al.*, 2021; Ding *et al.*, 2024; Zhou *et al.*, 2023). The greater biological impact of smaller particles is primarily due to their higher surface-to-volume ratio, which enhances reactivity and increases their potential for uptake by organisms (Koelmans *et al.*, 2022). Our findings reinforce the relevance of using predictive models to assess MP toxicity and improve our understanding of toxicity drivers.

A stronger focus on studies with well-characterized MP properties is essential to disentangle the contributions of individual properties to toxic outcomes (Liu *et al.*, 2024). Theoretically, ToMEx houses such comprehensive data, however, due to a lack of reporting, more detailed MP characterization (*e.g.*, surface charge and zeta potential) is unfortunately absent in over 80% of the data points. The consistent reporting of such properties will significantly advance both predictive modelling and regulatory frameworks. Furthermore, standardizing experimental designs in MP research, comparable to established frameworks in other fields (*e.g.*, the OECD guidelines for chemicals), would further enhance the comparability of results across studies (de Ruijter *et al.*, 2020). By identifying key properties linked to MP toxicity, models like those presented in this study can be used to recognize less hazardous property combinations and inform the design of safer plastic materials in the future.

The discrepancies between ROC-AUC and PR-AUC for the high quality dataset (random cross-validation = 0.02; blocked cross-validation = 0.10), compared to the imputed (random cross-validation = 0.07; blocked cross-validation = 0.17), complete-case (random cross-validation = 0.05; blocked cross-validation = 0.03), and low-quality datasets (random cross-validation = 0.04; blocked cross-validation = 0.12) suggest that model sensitivity to class imbalance (effect observed = ~35%; effect not observed = 65%) and to minority class-performance (*i.e.*, when toxic effect was observed) strongly depend on data quality and validation strategy. The superior performance of the high-quality dataset across

all evaluated metrics indicates that data quality, rather than class imbalance alone, is the primary driver for model performance.

Future studies could address class imbalance more directly using under- or oversampling techniques (*e.g.*, SMOTE). Another further extension would be the evaluation of cross-species generalization by implementing blocked cross-validation at the species level, allowing predictions for unobserved species during model training, similarly to the particle-level blocking applied in our study. Moving beyond binary effect classifications by incorporating effect sizes with associated uncertainty estimates would enable robust meta-analytical approaches to quantify the variance explained by experimental and biological factors and to predict effect-inducing concentrations (*e.g.*, LC50 values) and safety thresholds (*e.g.*, NOEC values). Ultimately, this would provide more actionable information for ecological risk assessment. Beyond effect size estimates and their associated uncertainties, databases must integrate data on associated contaminants, such as plastic additives (*e.g.*, by creating links to the PlastChem database; Monclús *et al.*, 2025), as well as detailed information on the dynamic environmental transformations (*e.g.*, aging or weathering) to which MP particles are subjected. Although outside the scope of the present study, including these variables would clearly provide a more accurate representation of the toxicological outcomes observed under more realistic conditions.

Within the current classification framework, our results indicate that the most crucial predictor groups for toxicity outcome prediction were the type of endpoint and the MP concentration. MP properties also contributed to predictive performance, despite the limited and frequently inconsistent information often reported. Given the current state of reported data in MP ecotoxicological research, treating MPs as a pollutant group in risk assessments, ideally using alignment methods (Koelmans *et al.*, 2020), might be an

appropriate approach at this stage. This approach could yield comparable risk estimates to support management decisions when associated uncertainties are quantitatively accounted (Coffin *et al.*, 2026). More detailed risk assessments incorporating other MP characteristics are likely to become increasingly valuable in the future, once a stronger mechanistic understanding and more comprehensive MP characterization become available.

Finally, our analyses are constrained to MP effects on aquatic organisms and cannot necessarily be extrapolated to other environmental compartments. As research on MP toxicity shifts towards a stronger focus on terrestrial ecosystems (de Souza Machado *et al.*, 2018; Holzinger *et al.*, 2022), particularly on MPs introduced into the environment through agricultural applications (Yang *et al.*, 2025), an extension of the ToMEx database or the development of similar databases on MP effects in terrestrial organisms would be highly valuable.

5. Conclusions

In toxicity prediction, focusing on data quality is more important than increasing the total volume of data. Consequently, prioritizing data quality is a promising strategy for reducing data mining effort while still improving predictive performance. Although MP properties are limitedly reported, a few MP properties, mainly particle length and density, emerge as important features for toxicity prediction. Greater standardization of experimental protocols, together with a more detailed MP characterization and higher reporting standards, could better support toxicity prediction, and therefore risk assessment frameworks. Such improvements will likely guide the development of safer, lower-toxicity materials. For any future application of the models developed in this study, the model evaluated using the high-quality dataset under blocked cross-validation should be used, as this approach prevents data leakage and provides a more realistic assessment of model

performance by accounting for potential dependencies between training and testing datasets.

Availability of data and materials

All data and code are openly available on GitHub (XX to be added upon acceptance XX) and Zenodo (XX to be added upon acceptance XX) [meanwhile the code and data are temporarily available here: <https://www.dropbox.com/scl/fo/2srlmrbzkwsg2ng8t0jcw/AIEzQvJ3LGrFN78k7vIt7sU?rlkey=4v1ijoh5ldz80oe90m2yu3jd3&st=1zzx2ea4&e=1&dl=0>]. Raw data were originally accessed via the Toxicity of Microplastics Explorer 2.0 database (<https://microplastics.sccwrp.org>).

References

- An D, Na J, Song J, Jung J. Size-dependent chronic toxicity of fragmented polyethylene microplastics to *Daphnia magna*. *Chemosphere*. 2021;271:129591. doi:10.1016/j.chemosphere.2021.129591
- Abarghouei S, Hedayati A, Raeisi M, Hadavand BS, Rezaei H, Abed-Elmdoust A. Size-dependent effects of microplastics on uptake, immune system, related gene expression, and histopathology of goldfish (*Carassius auratus*). *Chemosphere*. 2021;276:129977. doi:10.1016/j.chemosphere.2021.129977
- Başaran Kankılıç G, Koraltan İ, Erkmen B, Çağan AS, Çırak T, Özen M, Seyfe M, Altındağ A, Tavşanoğlu ÜN. Size-selective microplastic uptake by freshwater organisms: Fish, mussel, and zooplankton. *Environ Pollut*. 2023;336:122445. doi:10.1016/j.envpol.2023.122445

583 Choi D, Kim C, Kim T, Park K, Im J, Hong J. Potential threat of microplastics to
584 humans: Toxicity prediction modeling by small data analysis. *Environ Sci Nano*.
585 2023;10(4):1096–1108. doi:10.1039/D2EN00192F

586 Coffin S, Weisberg SB, Rochman C, Kooi M, Koelmans AA. Risk characterization of
587 microplastics in San Francisco Bay, California. *Microplast Nanoplast*. 2022;2:10.
588 doi:10.1186/s43591-022-00037-z

589 Coffin S, Bertrand L, Ahmed KT, Leite LDS, Cowger W, Siña M, et al. A probabilistic
590 risk framework for microplastics integrating uncertainty across toxicological and
591 environmental variability: development and application to marine and freshwater
592 ecosystems. *J Hazard Mater*. 2026;411:141021. doi:10.1016/j.jhazmat.2025.141021

593 Cui X, Yang T, Li Z, Nowack B. Meta-analysis of the hazards of microplastics in
594 freshwaters using species sensitivity distributions. *J Hazard Mater*. 2024;463:132919.
595 doi:10.1016/j.jhazmat.2023.132919

596 de Ruijter VN, Redondo-Hasselerharm PE, Gouin T, Koelmans AA. Quality criteria for
597 microplastic effect studies in the context of risk assessment: a critical review. *Environ Sci*
598 *Technol*. 2020;54(19):11692–11705. doi:10.1021/acs.est.0c03057

599 de Souza Machado AA, Kloas W, Zarfl C, Hempel S, Rillig MC. Microplastics as an
600 emerging threat to terrestrial ecosystems. *Glob Change Biol*. 2018;24:1405–1416.
601 doi:10.1111/gcb.14020

602 Devipriya K, Tlija M, Kumar CK, Kumar VC, Jana S, Jana C. Automated micro-plastic
603 detection and classification using deep convolution neural network pre-trained models
604 and transfer learning. *AIP Adv*. 2025;15:015301. doi:10.1063/5.0243976

Ding R, Chen Y, Shi X, Li Y, Yu Y, Sun Z, et al. Size-dependent toxicity of polystyrene microplastics on the gastrointestinal tract: oxidative stress related-DNA damage and potential carcinogenicity. *Sci Total Environ.* 2024;912:169514. doi:10.1016/j.scitotenv.2023.169514

Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. *J Anim Ecol.* 2008;77(4):802–813. doi:10.1111/j.1365-2656.2008.01390.x

European Chemicals Agency. New approach methodologies in regulatory science: proceedings of a scientific workshop: Helsinki, 19–20 April 2016. Publications Office of the European Union; 2016. doi:10.2823/543644

Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: learning a variable’s importance by studying an entire class of prediction models simultaneously. *J Mach Learn Res.* 2019;20:1–88.

Gong L, Martinez O, Mesquita P, Kurtz K, Xu Y, Lin Y. A microfluidic approach for label-free identification of small-sized microplastics in seawater. *Sci Rep.* 2023;13:12345. doi:10.1038/s41598-023-37900-9

Hartmann NB, Hüffer T, Thompson RC, Hassellöv M, Verschoor A, Dugaard AE, et al. Are we speaking the same language? Recommendations for a definition and categorization framework for plastic debris. *Environ Sci Technol.* 2019;53(3):1039–1047. doi:10.1021/acs.est.8b05297

Hodkovicova N, Hollerova A, Svobodova Z, Faldyna M, Faggio C. Effects of plastic particles on aquatic invertebrates and fish – a review. *Environ Toxicol Pharmacol.* 2022;96:104013. doi:10.1016/j.etap.2022.104013

627 Holzinger A, Mair MM, Lückner D, Seidenath D, Opel T, Langhof N, et al. Comparison of
628 fitness effects in the earthworm *Eisenia fetida* after exposure to single or multiple
629 anthropogenic pollutants. *Sci Total Environ.* 2022;838:156387.
630 doi:10.1016/j.scitotenv.2022.156387

631 Jin H, Kong F, Li X, Shen J. Artificial intelligence in microplastic detection and pollution
632 control. *Environ Res.* 2024;262:119812. doi:10.1016/j.envres.2024.119812

633 Kennedy SB, Antonio Vital AL, Kukkola A, Miller E, Yeh A, Coffin S, et al. Trends in
634 quality and risk assessment applicability of microplastic ecotoxicity studies. *J Hazard*
635 *Mater Adv.* 2026;21:100942. doi:10.1016/j.hazadv.2025.100942

636 Khanam MM, Uddin MK, Kazi JU. Advances in machine learning for the detection and
637 characterization of microplastics in the environment. *Front Environ Sci.*
638 2025;13:1573579. doi:10.3389/fenvs.2025.1573579

639 Kim SA, Kim L, Kim TH, An YJ. Assessing the size-dependent effects of microplastics
640 on zebrafish larvae through fish lateral line system and gut damage. *Mar Pollut Bull.*
641 2022;185:114279. doi:10.1016/j.marpolbul.2022.114279

642 Koelmans AA, Redondo-Hasselerharm PE, Mohamed Nor NH, Kooi M. Solving the
643 nonalignment of methods and approaches used in microplastic research to characterize
644 risk consistently. *Environ Sci Technol.* 2020;54(21):12307–12315.
645 doi:10.1021/acs.est.0c02982

646 Koelmans AA, Gebreyohanes Belay BM, Mintenig SM, Mohamed Nor NH, Redondo-
647 Hasselerharm PE, de Ruijter VN. Towards a rational and efficient risk assessment for
648 microplastics. *TrAC Trends Anal Chem.* 2023;165:117142.
649 doi:10.1016/j.trac.2023.117142

650 Koelmans AA, Redondo-Hasselerharm PE, Nor NHM, de Ruijter VN, Mintenig SM,
 651 Kooi M. Risk assessment of microplastic particles. *Nat Rev Mater.* 2022;7:138–152.
 652 doi:10.1038/s41578-021-00411-y

653 Lambert S, Scherer C, Wagner M. Ecotoxicity testing of microplastics: considering the
 654 heterogeneity of physicochemical properties. *Integr Environ Assess Manag.*
 655 2017;13(3):470–475. doi:10.1002/ieam.1901

656 Lei L, Wu S, Lu S, Liu M, Song Y, Fu Z, Shi H, Raley-Susman KM, He D. Microplastic
 657 particles cause intestinal damage and other adverse effects in zebrafish *Danio rerio* and
 658 nematode *Caenorhabditis elegans*. *Sci Total Environ.* 2018;619-620:1–8.
 659 doi:10.1016/j.scitotenv.2017.11.103

660 Li Z, Zhang Y, Chen Y. Toxicity of nanoplastics: machine learning combined with meta-
 661 analysis. *Nanoscale Horiz.* 2025;10:2017–2036. doi:10.1039/D5NH00024F

662 Liu C, Zong C, Chen S, Chu J, Yang Y, Pan Y, et al. Machine learning-driven QSAR
 663 models for predicting the cytotoxicity of five common microplastics. *Toxicology.*
 664 2024;508:153918. doi:10.1016/j.tox.2024.153918

665 Mayer M. missRanger: fast imputation of missing values (version 2.6.1) [computer
 666 software]. R package. 2024. [https://cran.r-](https://cran.r-project.org/web/packages/missRanger/index.html)
 667 [project.org/web/packages/missRanger/index.html](https://cran.r-project.org/web/packages/missRanger/index.html)

668 Mehinto AC, Coffin S, Koelmans AA, Brander SM, Wagner M, Thornton Hampton LM,
 669 et al. Risk-based management framework for microplastics in aquatic ecosystems.
 670 *Microplast Nanoplast.* 2022;2:12. doi:10.1186/s43591-022-00033-3

671 Menzel T, Meides N, Mauel A, Mansfeld U, Kretschmer W, Kuhn M, et al. Degradation
672 of low-density polyethylene to nanoplastic particles by accelerated weathering. *Sci Total*
673 *Environ.* 2022;826:154035. doi:10.1016/j.scitotenv.2022.154035

674 Meyers N, De Witte B, Schmidt N, Herzke D, Fuda JL, Vanavermaete D, et al. From
675 microplastics to pixels: testing the robustness of two machine learning approaches for
676 automated, Nile red-based marine microplastic identification. *Environ Sci Pollut Res.*
677 2024;31:61860–61875. doi:10.1007/s11356-024-35289-0

678 Mohammed S, Budach L, Feuerpfeil M, Ihde N, Nathansen A, Noack N, et al. The effects
679 of data quality on machine learning performance on tabular data. *Inf Syst.*
680 2025;132:102549. doi:10.1016/j.is.2025.102549

681 Monclús L, Arp HP, Groh KJ, Faltynkova A, Løseth ME, Muncke J, et al. Mapping the
682 chemical complexity of plastics. *Nature.* 2025;643:349–355. doi:10.1038/s41586-025-
683 09184-8

684 Müller S. How to crack a SMILES: automatic crosschecked chemical structure resolution
685 across multiple services using MoleculeResolver. *J Cheminform.* 2025;17:14.
686 doi:10.1186/s13321-025-01064-7

687 Pichler M, Hartig F. Can predictive models be used for causal inference? *arXiv.* 2023a.
688 doi:10.48550/arXiv.2306.10551

689 Pichler M, Hartig F. Machine learning and deep learning—A review for ecologists.
690 *Methods Ecol Evol.* 2023b;14:994–1016. doi:10.1111/2041-210X.14061

691 R Core Team. R: A language and environment for statistical computing. Vienna, Austria:
692 R Foundation for Statistical Computing; 2024.

693 Ramsperger AFRM, Jasinski J, Völkl M, Witzmann T, Meinhart M, Jérôme V, et al.
 694 Supposedly identical microplastic particles substantially differ in material properties,
 695 influencing particle-cell interactions and cellular responses. *J Hazard Mater.*
 696 2022;425:127961. doi:10.1016/j.jhazmat.2021.127961

697 Ramsperger AFRM, Narayana VKB, Gross W, Mohanraj J, Thelakkat M, Greiner A, et
 698 al. Environmental exposure enhances the internalization of microplastic particles into
 699 cells. *Sci Adv.* 2020;6(50):eabd1211. doi:10.1126/sciadv.abd1211

700 Renftle M, Trittenbach H, Poznic M, Heil R. What do algorithms explain? The issue of
 701 the goals and capabilities of explainable artificial intelligence (XAI). *Hum Soc Sci*
 702 *Commun.* 2024;11:1–10. doi:10.1057/s41599-024-03277-x

703 Rochman CM, Brookson C, Bikker J, Djuric N, Earn A, Bucci K, et al. Rethinking
 704 microplastics as a diverse contaminant suite. *Environ Toxicol Chem.* 2019;38(3):703–
 705 711. doi:10.1002/etc.4371

706 Rudin C. Stop explaining black-box machine learning models for high-stakes decisions
 707 and use interpretable models instead. *Nat Mach Intell.* 2019;1(5):206–215.
 708 doi:10.1038/s42256-019-0048-x

709 Scherer C, Brennholt N, Reifferscheid G, Wagner M. Feeding type and development
 710 drive the ingestion of microplastics by freshwater invertebrates. *Sci Rep.*
 711 2017;7(1):17006. doi:10.1038/s41598-017-17191-7

712 Schrank I, Trotter B, Dummert J, Scholz-Böttcher BM, Löder MGJ, Laforsch C. Effects
 713 of microplastic particles and leaching additive on the life history and morphology of
 714 *Daphnia magna*. *Environ Pollut.* 2019;255:113233. doi:10.1016/j.envpol.2019.113233

715 Schwarzer M, Brehm J, Vollmer M, Jasinski J, Xu C, Zainuddin S, et al. Shape, size, and
 716 polymer-dependent effects of microplastics on *Daphnia magna*. *J Hazard Mater*.
 717 2022;426:128136. doi:10.1016/j.jhazmat.2021.128136

718 Seewoo BJ, Goodes LM, Mofflin L, Mulders YR, Wong EV, Toshniwal P, et al. The
 719 plastic health map: a systematic evidence map of human health studies on plastic-
 720 associated chemicals. *Environ Int*. 2023;181:108225. doi:10.1016/j.envint.2023.108225

721 Tang F, Ishwaran H. Random Forest Missing Data Algorithms. *Stat Anal Data Min*.
 722 2017;10:363–377. doi:10.1002/sam.11348

723 Tang Y, Liu Y, Chen Y, Zhang W, Zhao J, He S, et al. A review: research progress on
 724 microplastic pollutants in aquatic environments. *Sci Total Environ*. 2021;766:142572.
 725 doi:10.1016/j.scitotenv.2020.142572

726 Thompson RC, Courteney-Jones W, Boucher J, Pahl S, Raubenheimer K, Koelmans AA.
 727 Twenty years of microplastics pollution research—What have we learned? *Science*.
 728 2024;383(6685):eadl2746. doi:10.1126/science.adl2746

729 Thornton Hampton LM, Lowman H, Coffin S, Darin E, De Frond H, Hermabessiere L, et
 730 al. A living tool for the continued exploration of microplastic toxicity. *Microplast*
 731 *Nanoplast*. 2022;2(1):13. doi:10.1186/s43591-022-00032-4

732 Thornton Hampton LM, Wyler DB, Almroth BC, Coffin S, Cowger W, Doyle D, et al.
 733 The Toxicity of Microplastics Explorer (ToMEx) 2.0. *Microplast Nanoplast*. 2025;5:12.
 734 doi:10.1186/s43591-025-00145-6

735 Tran HT, Hadi M, Nguyen TTH, Hoang HGG, Nguyen MK, Nguyen KN, et al. Machine
 736 learning approaches for predicting microplastic pollution in peatland areas. *Mar Pollut*
 737 *Bull.* 2023;194:115417. doi:10.1016/j.marpolbul.2023.115417

738 Tropsha A. Best practices for QSAR model development, validation, and exploitation.
 739 *Mol Inform.* 2010;29(6):476–488. doi:10.1002/minf.201000061

740 Wang M, Wang C, Yang X, Mu Y, Zhao X, Wu F, et al. Machine-learning-accelerated
 741 prediction of water quality criteria for microplastics. *ACS ES&T Water.*
 742 2026;acsestwater.5c00602. doi:10.1021/acsestwater.5c00602

743 Wieland S, Ramsperger AFRM, Gross W, Lehmann M, Witzmann T, Caspari A, et al.
 744 Nominally identical microplastic models differ greatly in their particle-cell interactions.
 745 *Nat Commun.* 2024;15(1):922. doi:10.1038/s41467-024-45281-4

746 Wieland S, Balmes A, Bender J, Kitzinger J, Meyer F, Ramsperger AF, et al. From
 747 properties to toxicity: comparing microplastics to other airborne microparticles. *J Hazard*
 748 *Mater.* 2022;428:128151. doi:10.1016/j.jhazmat.2021.128151

749 Wright SL, Thompson RC, Galloway TS. The physical impacts of microplastics on
 750 marine organisms: A review. *Environ Pollut.* 2013;178:483–492.
 751 doi:10.1016/j.envpol.2013.02.031

752 Yang L, Yang W, Li Q, Zhao Z, Zhou H, Wu P. Microplastics in Agricultural Soils:
 753 Sources, Fate, and Interactions with Other Contaminants. *J Agric Food Chem.*
 754 2025;73:12548–12562. doi:10.1021/acs.jafc.5c03682

755 Yu F, Hu X. Machine learning may accelerate the recognition and control of microplastic
756 pollution: future prospects. *J Hazard Mater.* 2022;432:128730.
757 doi:10.1016/j.jhazmat.2022.128730

758 Yu Y, Kumar M, Bolan S, Padhye LP, Bolan N, Li S, et al. Various additive release from
759 microplastics and their toxicity in aquatic environments. *Environ Pollut.*
760 2024;343:123219. doi:10.1016/j.envpol.2023.123219

761 Zhang J, Li X, Chen H, Liu G, Yan X, Yan B. Unraveling the ecotoxicity of
762 micro(nano)plastics loaded with environmental pollutants using ensemble machine
763 learning. *J Hazard Mater.* 2025;495:138911. doi:10.1016/j.jhazmat.2025.138911

764 Zhen Y, Wang L, Sun H, Liu C. Prediction of microplastic abundance in surface water of
765 the ocean and influencing factors based on ensemble learning. *Environ Pollut.*
766 2023;331:121834. doi:10.1016/j.envpol.2023.121834

767 Zhou N, Wang Z, Yang L, Zhou W, Qin Z, Zhang H. Size-dependent toxicological
768 effects of polystyrene microplastics in the shrimp *Litopenaeus vannamei* using a
769 histomorphology, microbiome, and metabolic approach. *Environ Pollut.*
770 2023;316:120635. doi:10.1016/j.envpol.2022.120635

771

773 **Table S1.** Number of data points for the training, validation, and test subsets of the four datasets
774 used under random cross-validation.

IMPUTED		
	Number of data points with effect absent (response = 0)	Number of data points with effect present (response = 1)
Training	4602	2349
Validation	1546	793
Test	1513	787
COMPLETE-CASE		
	Number of data points with effect absent (response = 0)	Number of data points with effect present (response = 1)
Training	1557	1150
Validation	519	383
Test	520	385
HIGH-QUALITY		
	Number of data points with effect absent (response = 0)	Number of data points with effect present (response = 1)
Training	687	535
Validation	229	178
Test	230	179
LOW-QUALITY		
	Number of data points with effect absent (response = 0)	Number of data points with effect present (response = 1)
Training	639	450
Validation	213	150
Test	214	150

775
776

777
778 **Table S2.** Number of particle identifiers (particle ID) and data points for the training, validation,
779 and test subsets of the four datasets used under blocked cross-validation.

IMPUTED			
	Number of particle IDs	Number of data points with effect absent (response = 0)	Number of data points with effect present (response = 1)
Training	309	4497	2439
Validation	105	1390	673
Test	101	1676	729
COMPLETE-CASE			
	Number of particle IDs	Number of data points with effect absent (response = 0)	Number of data points with effect present (response = 1)
Training	119	1362	895
Validation	42	622	485
Test	39	612	538
HIGH-QUALITY			
	Number of particle IDs	Number of data points with effect absent (response = 0)	Number of data points with effect present (response = 1)
Training	53	681	488
Validation	19	332	311
Test	17	133	93
LOW-QUALITY			
	Number of particle IDs	Number of data points with effect absent (response = 0)	Number of data points with effect present (response = 1)
Training	77	718	535
Validation	27	190	122
Test	25	158	93

780

781

782 **Table S3.** List of hyperparameter ranges used for model tuning under a 100-hyperparameter grid.

Hyperparameter	Value
max_depth	from = 4; to = 20; by = 1
eta	from = 0.01; to = 0.1; by = 0.01
n.rounds	from = 100; to = 700: by = 30
lambda	0, 1, 2.5, or 5
alpha	0, 0.5, 1, or 5

783

784 **Table S4.** Area under the curve (AUC) values for training, validation, and test sets across the
785 four datasets using random 5-fold cross-validation (CV).

	Imputed	Complete-case	High-quality	Low-quality
Training (5-fold CV average)	AUC = 0.87	AUC = 0.86	AUC = 0.87	AUC = 0.83
Validation	AUC = 0.87	AUC = 0.87	AUC = 0.89	AUC = 0.89
Test	AUC = 0.89	AUC = 0.87	AUC = 0.93	AUC = 0.83

786

787 **Table S5.** Area under the curve (AUC) values for training, validation, and test sets across the
788 four datasets using blocked 5-fold cross-validation (CV) by particle identifier.

	Imputed	Complete-case	High-quality	Low-quality
Training (5-fold CV average)	AUC = 0.74	AUC = 0.73	AUC = 0.74	AUC = 0.67
Validation	AUC = 0.70	AUC = 0.74	AUC = 0.66	AUC = 0.45
Test	AUC = 0.69	AUC = 0.67	AUC = 0.84	AUC = 0.75

789

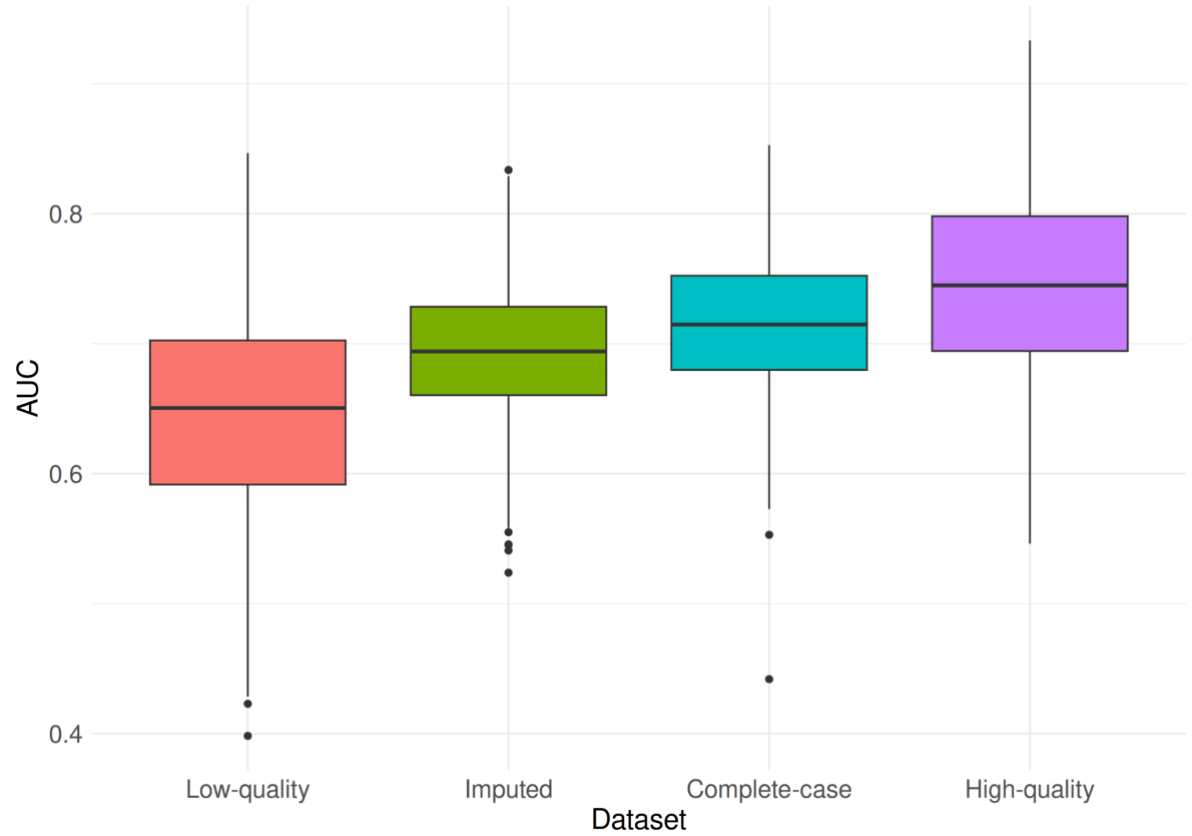
790

791

792

793 **Table S6.** Matthew's Correlation Coefficient (MCC) and sensitivity for the best area under the
794 curve (AUC) threshold under the random and blocked cross-validation schemes.

Cross-validation	Dataset	Threshold	Sensitivity	MCC
Random	Imputed	0.292	0.848	0.596
	Complete-case	0.355	0.841	0.564
	High-quality	0.515	0.815	0.691
	Low-quality	0.549	0.680	0.541
Blocked	Imputed	0.411	0.492	0.277
	Complete-case	0.408	0.589	0.275
	High-quality	0.175	0.924	0.584
	Low-quality	0.425	0.784	0.398



795

796 **Figure S1.** Comparison of model performance across standardized dataset scenarios. The
797 boxplots represent the distribution of Area Under the Curve (AUC) values obtained over 1,000
798 random iterations for each dataset downsampled to a fixed size of 1,816 data points and to only
799 common variables after one-hot-encoding.

800 **Table S7.** Feature-group ablation analysis for the best model using the high-quality dataset and
801 random cross-validation scheme.

mean_auc	sd_auc	Model
0.908011	0.026373	Concentration_Endpoints_Experiment_Microplastic_Species
0.906973	0.019077	Concentration_Endpoints_Experiment_Microplastic
0.900483	0.014168	Concentration_Endpoints_Microplastic_Species
0.897651	0.020335	Concentration_Endpoints_Microplastic
0.896738	0.022341	Concentration_Endpoints_Experiment_Species
0.894757	0.021072	Concentration_Endpoints_Experiment
0.891477	0.019793	Concentration_Endpoints_Species
0.871892	0.023248	Endpoints_Experiment_Microplastic_Species
0.869044	0.027712	Endpoints_Experiment_Microplastic
0.86838	0.021949	Endpoints_Microplastic_Species
0.866453	0.021743	Endpoints_Microplastic
0.864832	0.033129	Concentration_Endpoints
0.856516	0.035559	Endpoints_Experiment
0.848123	0.013785	Endpoints_Species
0.84791	0.031348	Endpoints_Experiment_Species
0.844687	0.022734	Concentration_Experiment_Microplastic
0.843273	0.029726	Concentration_Experiment_Microplastic_Species
0.841759	0.027604	Concentration_Experiment_Species
0.839635	0.029983	Concentration_Experiment

0.832956	0.031243	Concentration_Microplastic_Species
0.822586	0.040657	Concentration_Species
0.815285	0.03257	Endpoints
0.813982	0.032823	Experiment_Microplastic_Species
0.81305	0.031558	Concentration_Microplastic
0.809664	0.022809	Experiment_Microplastic
0.796443	0.027439	Microplastic_Species
0.793468	0.031054	Experiment
0.792413	0.032037	Experiment_Species
0.772251	0.03167	Species
0.759661	0.029808	Concentration
0.75925	0.042887	Microplastic

803
804
805 **Table S8.** Feature-group ablation analysis for the best model using the high-quality dataset and
806 blocked cross-validation scheme.
807

mean_auc	sd_auc	Model
0.791194	0.106198	Concentration_Endpoints_Experiment_Microplastic_Species
0.772453	0.088537	Concentration_Endpoints_Experiment_Microplastic
0.762452	0.092881	Concentration_Endpoints_Experiment_Species
0.754931	0.100117	Concentration_Endpoints_Experiment
0.753492	0.124127	Endpoints
0.75325	0.136207	Endpoints_Species
0.751207	0.070851	Concentration_Experiment_Species
0.748724	0.081471	Concentration_Endpoints_Microplastic_Species
0.748592	0.095408	Concentration_Endpoints_Species
0.745459	0.073797	Endpoints_Experiment_Microplastic_Species
0.735954	0.129334	Concentration_Experiment_Microplastic_Species
0.732257	0.105222	Concentration_Endpoints
0.731459	0.101135	Concentration_Endpoints_Microplastic
0.731072	0.089249	Endpoints_Experiment_Species
0.723161	0.111999	Endpoints_Microplastic

0.715774	0.116305	Endpoints_Microplastic_Species
0.713725	0.1453	Endpoints_Experiment
0.695504	0.112843	Experiment_Microplastic
0.693822	0.129767	Experiment
0.693664	0.085253	Concentration_Experiment_Microplastic
0.691553	0.089966	Endpoints_Experiment_Microplastic
0.688543	0.092642	Species
0.685769	0.098804	Concentration_Experiment
0.684348	0.12086	Concentration_Microplastic_Species
0.670804	0.06134	Concentration_Species
0.655506	0.166506	Experiment_Species
0.647225	0.148763	Experiment_Microplastic_Species
0.639003	0.150935	Microplastic_Species
0.614335	0.143571	Microplastic
0.578145	0.044314	Concentration
0.573492	0.103594	Concentration_Microplastic

808

809