High Data Quality Enhances Microplastic Toxicity Prediction

Ana L. Antonio Vital[1], Scott Coffin[2], Andrea Bonisoli-Alquati[3], Maaike Vercauteren[4,] Luan de Souza Leite[5], Maximilian Pichler[6], Magdalena M. Mair[1,7*]

[1]Statistical Ecotoxicology, University of Bayreuth, Bayreuth, Germany

[2]California State Water Resources Control Board, Sacramento, USA

[3]California State Polytechnic University, Pomona, USA

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

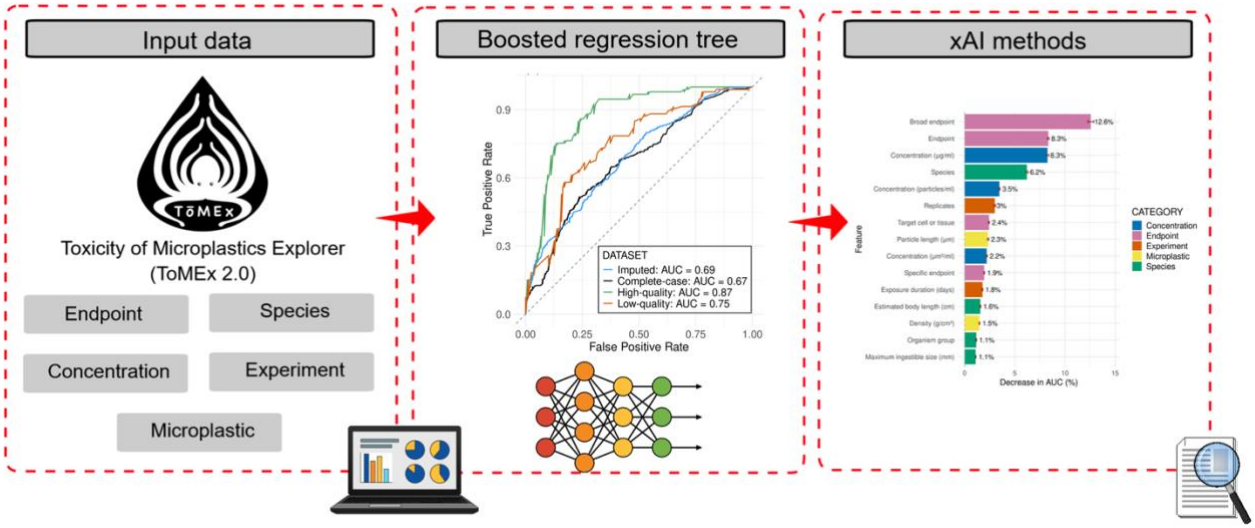
[5]State University of Campinas (UNICAMP), Campinas, Brazil

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

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

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

Graphical Abstract



Abstract

Unlike chemicals, microplastics (MPs) lack standardized identifiers, limiting the applicability of traditional predictive ecotoxicology methods such as quantitative structure-activity relationship (QSAR) models. This study aimed to predict MP toxicity using MP properties, MP concentration, organismal traits, endpoints, and experimental design, and to evaluate how data pre-processing, dataset size, and quality influence model performance. 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 for both random cross-validation (AUC = 0.93) and blocked cross-validation by particle identifier (AUC = 0.87). Explainable artificial intelligence (xAI) analyses showed that predictive performance was primarily determined by endpoints and concentration, with MP properties contributing despite limited reporting. Our findings demonstrate the feasibility of machine learning to predict and identify key drivers of MP toxicity, highlighting that high-quality data improves predictive performance while reducing data mining and computational costs. Standardized experiments, detailed MP characterization, and high reporting standards would better support risk assessment frameworks and inform the design of safer materials.

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

1. Introduction

Microplastics (MP; < 5 mm) are a complex and multidimensional group of pollutants (Rochman *et al.*, 2019). 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 comprehensively 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 or 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*, *in-silico*, and *in-vitro* approaches (European Chemicals Agency, 2016). NAMs aim to reduce animal testing and enable more cost- and time-efficient chemical assessments. Chemicals can

be linked to unique identifiers (*e.g.*, SMILES and CAS numbers) through their well-defined molecular structures and stable properties, which facilitates the use of *in-silico* NAMs by ensuring the integrity and reproducibility of chemical information across databases and publications (Müller, 2025). Within the NAM frameworks, particularly quantitative structure-activity relationship (QSAR) modeling, has demonstrated robust predictive capabilities for chemicals by establishing mathematical correlations between chemical structure and its toxic outcomes (Tropsha, 2010).

Within QSAR modeling, particularly machine learning (ML) methods have emerged as a promising alternative to more rigid statistical models. More generally, ML techniques employ a variety of algorithms to identify patterns in data and generate predictions (Pichler & Hartig, 2023), establishing complex relationships between molecular structures and toxicological outcomes. First approaches have been published that apply ML-based modeling to predict MP toxicity by linking MP properties to observed effects (Wang *et al.*, 2026; Zhang *et al.*, 2025). The integration of plastic material properties into these models can support the development of safer, more environmentally friendly plastic materials, thereby further reducing the need for animal testing in material development. Furthermore, a predictive approach to MP toxicity can identify data gaps, highlighting areas that need further investigation. However, a limitation of deploying ML algorithms is their lack of interpretability (Rudin, 2009). Unlike statistical models, ML algorithms do not report coefficients that precisely address how traits are related to toxicity. Fortunately, explainable artificial intelligence (xAI) tools enable understanding of how and why ML algorithms make their predictions (Renftle *et al.*, 2024), potentially allowing identification of MP properties or experimental conditions that drive toxicity.

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 e*t 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). From toxicity studies, the focus is often on human health, with predictions based on tissue or cell cultures from limited datasets (*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 (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.

To increase the environmental relevance of these predictive modeling approaches for MP toxicity, larger, 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). For this task, almost 60 scientists worldwide mined and validated 6,927 data points from 124 publications, covering approximately two years of research on MP effects on aquatic organisms.

With the growing number of publications from MP research, there have been increasing calls to standardize key aspects of MP toxicity studies, which is also pivotal for QSAR of 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 thresholds, the ToMEx 2.0 database incorporates data scores based on de Ruijter *et al.* (2020) and Mehinto *et al*. (2022), rating each experiment according to the minimum technical and risk-assessment quality criteria to be deemed reliable and informative for hazard assessment (Kennedy *et al.*, 2025). However, the importance and effectiveness of data quality in predictive models of MP toxicity are still poorly understood.

In this study, we integrate information from 28 polymer types and 45 features related to MP properties, species, endpoint, and experimental design. We predicted MP toxicity and utilized xAI approaches to identify the most important features for toxicity prediction and thus identify key drivers of MP toxicity. For this task, and to understand how dataset size, data quality, and data pre-processing approach influence the predictions, we used four different data sets derived from the ToMEx 2.0: (*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 hypothesize 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 is likely to yield the best predictive performance.

2. Materials and Methods

Data source

As the basis for our predictive models, we used 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). In ToMEx 2.0, toxicity is defined as a binary response variable: toxic effect found (yes = 1) or not found (no = 0), based on whether the treatment containing MP particles differed statistically from the particle-free control as reported by the authors.

Data preparation

To increase taxonomic consistency in the dataset, we removed the data points that used plants and the unsuitable features from all four datasets (see File S2). In ToMEx 2.0, we often have several data points corresponding 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 blocked cross-validation.

Where possible, missing values were first manually imputed by selecting the most reasonable estimates. Missing temperature values were first imputed 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 for each organism

group (*i.e.*, higher taxonomic groups, *e.g.*, crustacea, fish, insects) Particles with lengths ≥ 1000 µm 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), we standardized the sample size 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/cm³) 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, when 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), related to quality scores (*e.g.*, particle size score, polymer type score), as well as features with above 51% of missing values (*e.g.*, pH, charge, zeta potential), were dropped. A lower threshold would lead to a drop in many relevant features, such as particle width and concentrations. For a complete list of features used and dropped, along with their corresponding NA 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 test and validation sets), 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 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, while blocked cross-validation assumes that data points are grouped (*e.g.*, by particle) and that the entire group should be assembled in the same cross-validation folder, in order to reflect predictions for unseen groups. Therefore, blocked cross-validation provides a more realistic scenario when the goal is to generalize new particles. 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) maintained an approximate ratio of 60% training, 20% validation, and 20% test concerning the whole dataset. 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) to avoid identical particles being present in more than one data subset. For the blocked cross-validation, each dataset (*i.e.*, imputed, complete-case, high-quality, and low-quality) also maintained an approximate 60% training, 20% validation, and 20% test split, based on the total number of particle IDs (see Table S2).

Modeling and evaluation

To predict the response variable (*i.e.*, MP toxicity), we used the decision-tree-based Boosted Regression Tree (BRT) method, an efficient ML algorithm suitable for both regression and classification tasks that improves predictive performance by sequentially fitting several models in an iterative process (Elith *et al.*, 2008). In ML research, it is widely accepted that classical ML algorithms, such as BRT (in particular, extreme boosting), achieve higher accuracy on tabular datasets. The performance of ML algorithms depends strongly on their

hyperparameters that, for example, control their complexity (*i.e.*, regularization). We performed hyperparameter tuning using a random grid of 100 parameter set combinations, first selected from hyperparameter ranges informed by our prior knowledge (see the list of parameters in Table S3).

Model performance was then evaluated using a 5-fold cross-validation on the training subset. In each iteration, one subset served as the validation set while the remaining four were used for training. For the blocked cross-validation by particle ID, the 5-fold split was based on the particle ID. We created the five folds with roughly equal numbers of particles, ensuring that all particles in each fold remained within that fold. After splitting, the particle IDs were removed and not used as a predictive feature. The average training performance was calculated for each of the five iterations under both cross-validation strategies. This process was repeated for each of the 100 hyperparameter combinations. The best-performing hyperparameter set was used to fit a BRT 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 models (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 present (yes = 1) or absent (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 a meaningful metric for binary classification, especially 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. 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 a high-quality dataset using the best-performing model under either random or blocked cross-validation. 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 original 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 across studies, 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 is varied 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 (finite difference). These local effects are then averaged across observations to obtain an average conditional effect for each predictor (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 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 performed well, demonstrating 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 = 0.90), complete-case (AUC = 0.87), and low-quality dataset (AUC = 0.84). Additional AUC values from the training phase are provided in Tables S4 and S5. The blocked cross-validation (Figure 1B) also performed best with the high-quality dataset (AUC = 0.87), followed by the low-quality dataset (AUC = 0.75), the imputed dataset (AUC = 0.69), and the complete-case dataset (AUC = 0.67). Under

random cross-validation (Figure 1C), PR-AUC values followed the order: high-quality (PR-AUC = 0.91), imputed (PR-AUC = 0.83), complete-case (PR-AUC = 0.82), and low-quality (PR-AUC = 0.78). Under blocked cross-validation (Figure 1D), PR-AUC values were highest for the high-quality dataset (PR-AUC = 0.78), followed by complete-case (PR-AUC = 0.64), low-quality (PR-AUC = 0.63), and imputed (PR-AUC = 0.51). Additional MCC values are reported in Table S6.

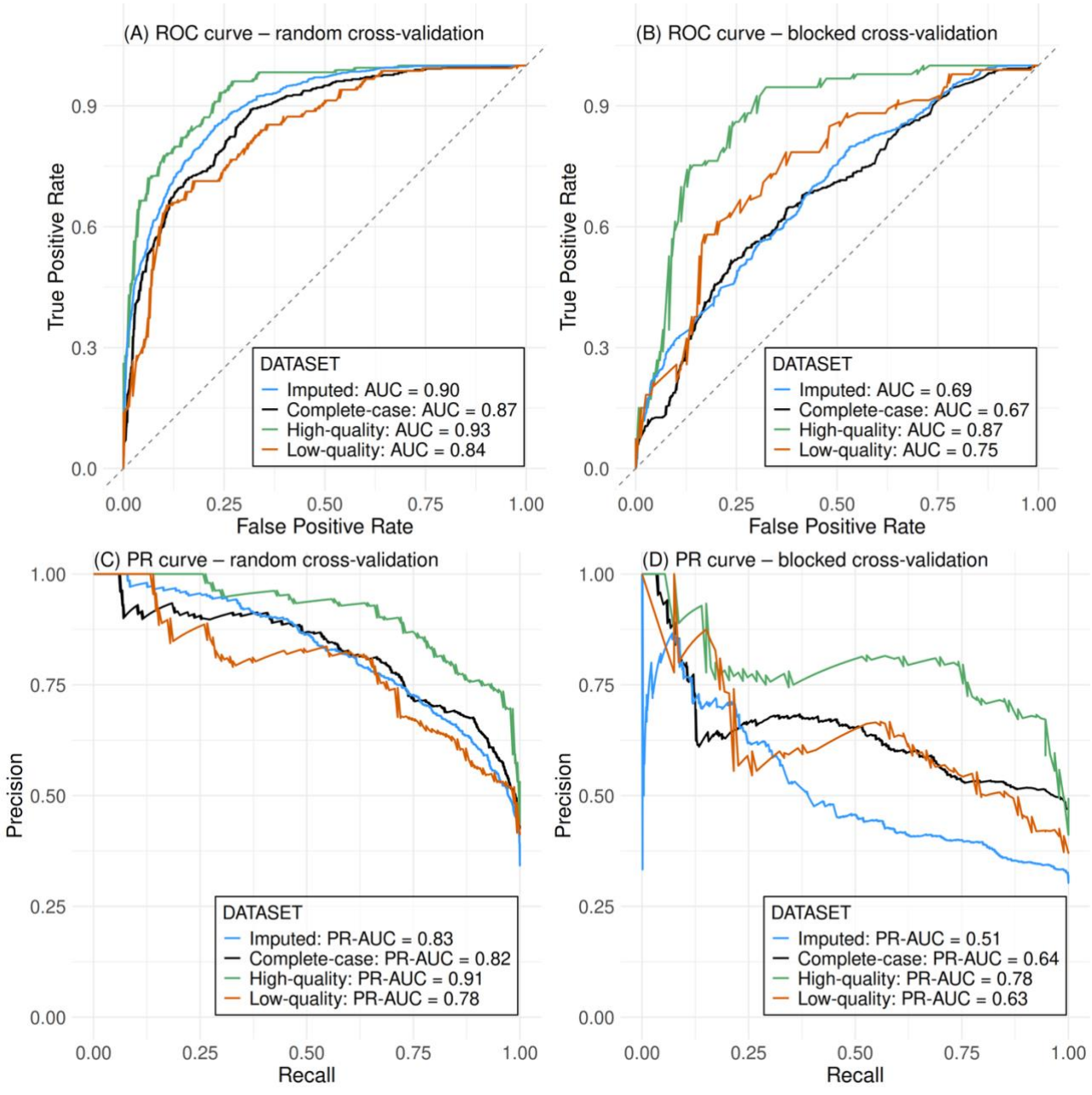


Figure 1. Comparison of the performance of models trained on imputed, complete-case, high-quality, and low-quality datasets. (A) and (B) show ROC (receiver operating characteristic) curves and AUC (area under the curve) values derived from predictions on the independent test subset for both cross-validation 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 calculation of permutation importance for the top 15 features (Figure 2) revealed that the broad endpoint (*e.g.*, fitness, metabolism, etc.) was the most crucial feature, resulting in an AUC decrease of 12.6%, followed by the endpoint (*e.g.*, population size, catalase activity; AUC decrease of 8.3%). Still within the endpoint category, the target cell or tissue (AUC decrease of 2.4%) and the specific endpoint (*e.g.*, reproduction, oxidative stress; AUC decrease of 1.9%) are among the top-ranked features. The three concentrations reported in different units also appeared: concentration in mass (µg/mL; AUC decrease = 8.3%), concentration in number of particles (particles/mL; AUC decrease = 3.5%), and concentration in volume (µm³/mL; AUC decrease = 2.2%). For the experiment category, replicates (AUC decrease = 3%) and exposure duration in days (AUC decrease = 1.8%) are also observed. From the species category, species (AUC decrease = 6.2%), estimated body length (cm; AUC decrease = 1.6%), organism group (AUC decrease = 1.1%), and maximum ingestible size (mm; AUC decrease = 1.1%) also contributed to the predictive performance. Two microplastic-related features: particle length (µm; AUC decrease = 2.3%) and density (g/cm^3; AUC decrease = 1.5%) complete the top 15 ranking of most important features.

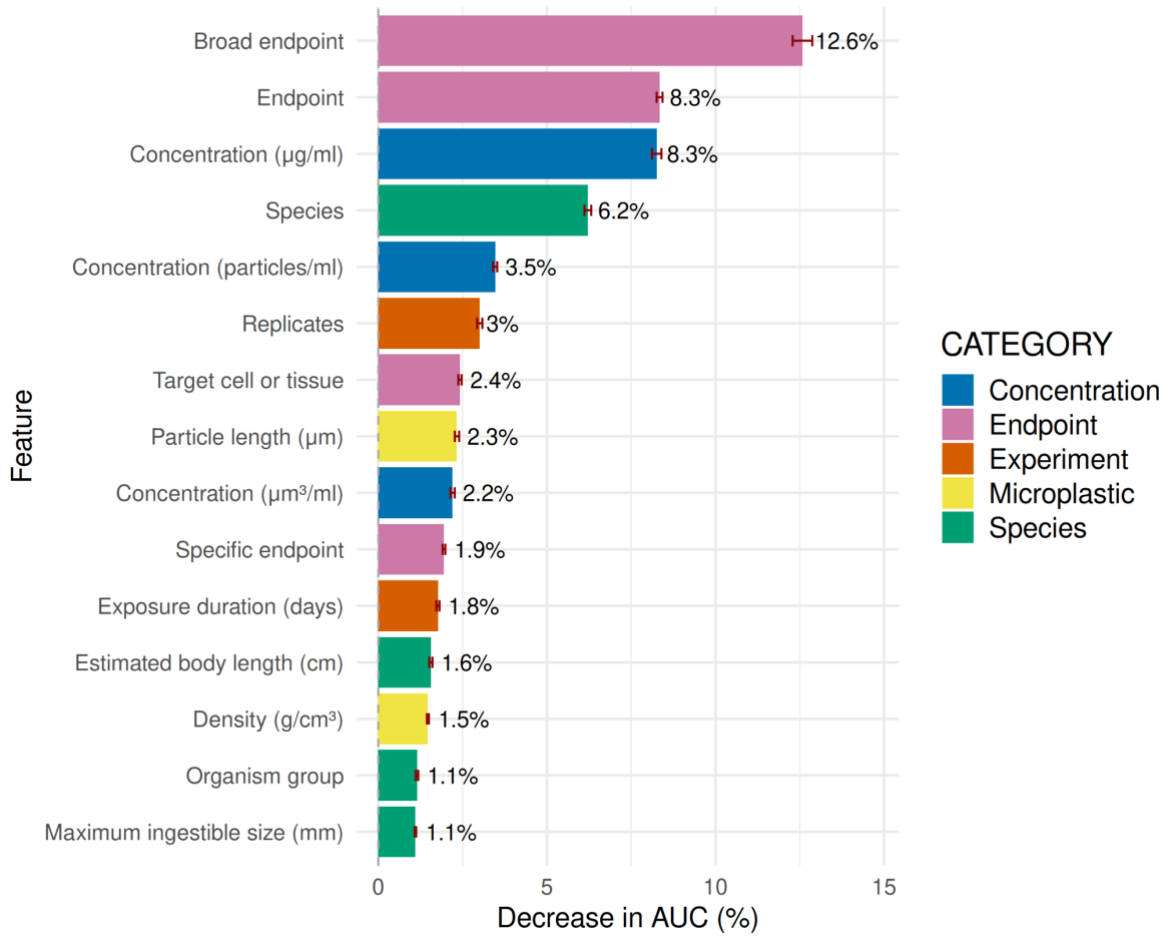


Figure 2. Top 15 decreases in AUC (area under the curve) relative to the baseline model after 20 random shuffles per feature for the final model on the high-quality dataset under the blocked cross-validation scheme. Colors indicate assignments of features to feature categories: concentration (blue), 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 (µm³/mL) and mass concentration (µg/mL). In contrast, the strongest positive interaction was observed between concentration in mass (µg/mL) and particle length (µm). The particle length (µm) also showed a strong positive interaction with the estimated body length (cm) and a weaker, but still positive, interaction with the estimated maximum ingestible size (mm).

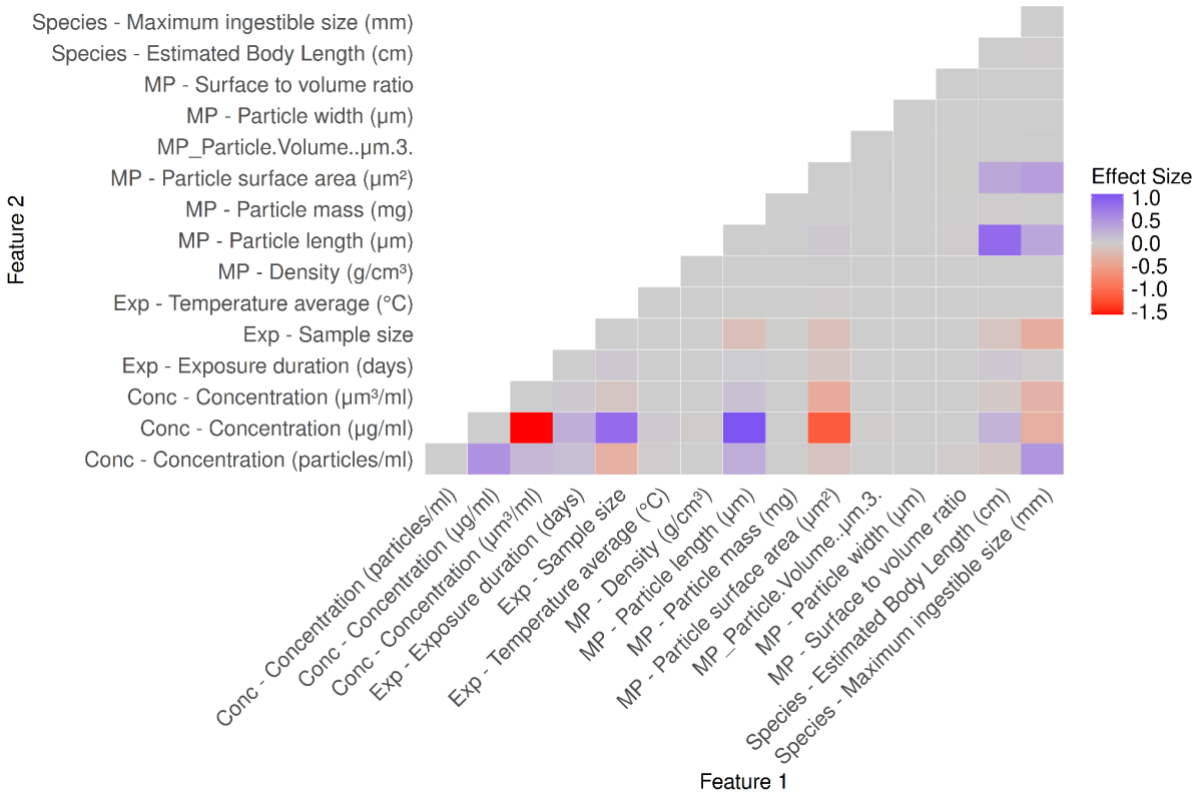


Figure 3. Heat map showing the pairwise interaction strengths of the average conditional effect (ACE) for the high-quality dataset fitted with the hyperparameters from the blocked cross-validation scheme. Interactions in blue increase, and those in red decrease, the probability of observing a toxicity 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 ± 0.025), followed by models with four categories, excluding either species (0.907 ± 0.019) or experiment (0.901 ± 0.014). For ablation analysis on single-category models, endpoint presented the highest performance (0.815 ± 0.033), followed by experiment (0.793 ± 0.031), species (0.772 ± 0.031), concentration (0.760 ± 0.030), and MP properties (0.759 ± 0.043). 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 four categories: concentration, endpoint, experiment, and species (0.782 ± 0.098); followed by the groups with concentration, endpoints, and experiment (0.779 ± 0.063), and concentration, endpoints, experiment, and MP (0.776 ± 0.096). Among the single-category models, endpoints ranked highest (0.727 ± 0.069), succeeded by species (0.698 ± 0.101), experiment (0.638 ± 0.106), MP (0.604 ± 0.120), and concentration (0.566 ± 0.084).

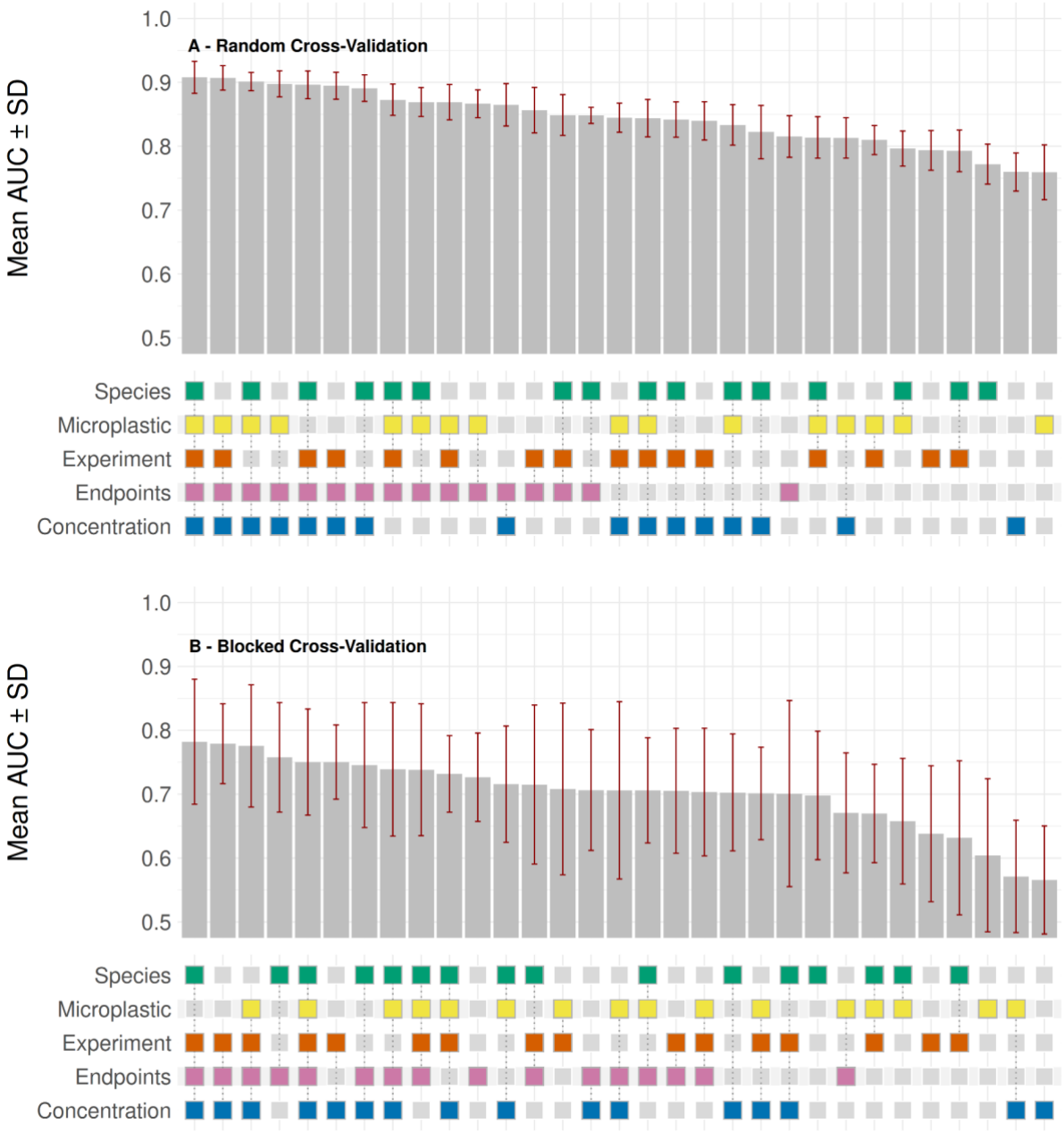


Figure 4. Barplot showing the mean AUC ± standard deviation for the high-quality dataset across models with different feature categories included and evaluated using 10-fold cross-validation. A: Random cross-validation. B: Blocked cross-validation. The inclusion of distinct feature sets in each model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

Discussion

Despite the high heterogeneity of our data, our models successfully predicted MP toxicity based on the data provided in ToMEx 2.0. Despite being one of the two smallest

datasets (~2,000 data points), the high-quality dataset outperformed the larger imputed (>11,000 data points) and complete-case (>4,000 data points) datasets. When comparing the two cross-validation schemes, the random scheme (AUC = 0.930) outperformed the blocked scheme (AUC = 0.870). This decline in performance under blocked cross-validation was expected, as the model had to predict the toxicity of new MP particles that were entirely unknown during training. In contrast, random cross-validation allows both training and testing on identical MP particles when several data points are derived from the same experiment, likely leading to data leakage. The practical use of predictive modeling approaches in ecotoxicology requires reliable toxicity predictions of entirely untested substances or materials, and this concept aligns with our blocked cross-validation scheme.

Beyond the cross-validation scheme, the effectiveness of predictive models is influenced by data quality, including bias, noise, sample size, and mislabeling (Mohammed *et al.*, 2025). Similarly, we found the best predictive performance on the high-quality dataset (~2,000 data points across 54 studies), demonstrating that focusing on data quality is more important than increasing the total amount of data. In consequence, prioritizing high-quality, well-characterized experiments and good data reporting is crucial to improve predictive performance while reducing data mining efforts and computational costs (Kennedy *et al.*, 2025).

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 (Figure 2 and 4). Based on current availability data from published studies, the association between toxic outcomes and MPs is 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 (Schrank *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 this limitation on MP research, a few studies have shown that meaningful toxicity predictions can still be derived from a small set of 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). The representative study conducted by Wang *et al.* (2025) used random forest to predict NOEC values using the ToMEx database as the data source and including only seven features (i.e., polymer type, shape, length, density, species group, endpoint, and environment) and found that MP length was the most important feature. Consistent with the findings of Zhang *et al.* (2025), particle size, measured as diameter, was 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 different 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 their reactivity and their potential for uptake by organisms (Koelmans *et al.*, 2022). These findings reinforce the relevance of using predictive models to assess MP toxicity and to improve our understanding of toxicity drivers.

To advance predictive modeling and regulatory frameworks, a stronger focus on characterizing MP properties is essential to disentangle their individual contributions to toxicity (Liu *et al.*, 2024). The standardization of experimental design in MP research, as 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 the key toxicological MP properties, models like those presented in our study could be developed to recognize less hazardous combinations of MP properties, thereby informing the design of safer plastic materials in the future.

The lower discrepancies between ROC-AUC and PR-AUC for the high quality dataset (random cross-validation = 0.02; blocked cross-validation = 0.09) compared to the imputed (random cross-validation = 0.07; blocked cross-validation = 0.18) and low-quality datasets (random cross-validation = 0.06; blocked cross-validation = 0.12) suggest that model sensitivity to class imbalance (effect yes = ~35%; effect no = 65%) and to minority class-performance (*i.e.*, when toxic effect occurs) strongly depend on data quality and validation strategy. The consistent 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 to evaluate cross-species generalization by implementing blocked cross-validation at the species level, allowing predictions for unseen species during model training, similarly to the particle-level blocking applied in our study. Additionally, extending the approach to regression-based models would enable the prediction of concentrations at which effects are expected to occur, providing more quantitative information for risk assessment. Future developments of these analyses, the ToMEx database, or any other effort in evidence synthesis, should quantify the documented effects of MP, rather than relying on a binary classification (*i.e.,* effect present or absent).

Within the current classification framework, our results indicate that the most crucial predictor groups for toxicity prediction were endpoints and concentration. MP properties have also contributed to predictive performance, despite the limited and frequently inconsistent

information currently reported. Given these current data limitations, treating MPs as a pollutant group in risk assessments, ideally using alignment methods (Koelmans *et al.*, 2020), may be an appropriate approach at this stage and can yield comparable risk estimates for management decisions when associated uncertainties are quantitatively accounted for (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.

Our findings suggest that, in toxicity prediction, focusing on data quality is more important than increasing the total volume of data. Prioritizing high-quality data therefore is a promising strategy to reduce 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 experiments ensuring more detailed MP characterization and higher reporting standards, could therefore better support risk assessment frameworks and guide the development of safer, lower-toxicity materials in the future.

Acknowledgements

The authors would like to thank Leah Thornton Hampton for her collaboration and leadership in developing the ToMEx database and Win Cowger for giving constructive feedback and valuable suggestions on the study.

Funding

ALAV and MMM received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1357 – 391977956. LdeSL received funding support from São Paulo Research Foundation (Proc. FAPESP 2023/16350–2; 2022/12104–4).

Credit statement

Ana L. Antonio Vital: Conceptualization, Methodology, Formal analysis, Data curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration. **Scott Coffin:** Conceptualization, Writing - Review & Editing. **Andrea Bonisoli-Alquati:** Conceptualization, Methodology, Writing - Review & Editing; **Maaike Vercauteren:** Conceptualization, Formal analysis, Writing - Review & Editing. **Luan de Souza Leite:** Conceptualization, Writing - Review & Editing, Visualization. **Maximilian Pichler:** Conceptualization, Methodology, Writing - Review & Editing. **Magdalena M. Mair:** Conceptualization, Methodology, Formal analysis, Writing - Review & Editing, Project administration, Funding acquisition.

Data statement

All data and code will be made publicly available upon publication. Raw data were originally accessed via the Toxicity of Microplastics Explorer 2.0 database (https://sccwrp.shinyapps.io/aq_mp_tox_shiny/).

Disclaimer

SC was employed at the California Environmental Protection Agency's Office of Environmental Health Hazard Assessment (OEHHA) and the State Water Resources Control Board (SWRCB) during the preparation of this manuscript. The views expressed are his own and do not necessarily reflect the opinions or policies of OEHHA, the SWRCB, or the California Environmental Protection Agency.

Statement of competing interests

There are no competing interests to declare.

References

An, D., Na, J., Song, J., Jung, J., 2021. Size-dependent chronic toxicity of fragmented polyethylene microplastics to Daphnia magna. Chemosphere 271, 129591. https://doi.org/10.1016/j.chemosphere.2021.129591

Abarghouei, S., Hedayati, A., Raeisi, M., Hadavand, B. S., Rezaei, H., & Abed-Elmdoust, A. 2021. Size-dependent effects of microplastics on uptake, immune system, related gene expression, and histopathology of goldfish (Carassius auratus). Chemosphere, 276, 129977. https://doi.org/10.1016/j.chemosphere.2021.129977

Choi, D., Kim, C., Kim, T., Park, K., Im, J., & Hong, J. 2023. Potential threat of microplastics to humans: Toxicity prediction modeling by small data analysis. Environmental Science: Nano, 10(4), 1096–1108. https://doi.org/10.1039/D2EN00192F

Coffin, S., Weisberg, S.B., Rochman, C., Kooi, M., Koelmans, A.A., 2022. Risk characterization of microplastics in San Francisco Bay, California. Microplastics and Nanoplastics 2. https://doi.org/10.1186/s43591-022-00037-z

Coffin, S., Bertrand, L., Ahmed, K.T., Leite, L.D.S., Cowger, W., Siña, M., Barrick, A., Kukkola, A., Almroth, B.C., Miller, E., Yeh, A., 2026. A Probabilistic Risk Framework for Microplastics Integrating Uncertainty Across Toxicological and Environmental Variability: Development and Application to Marine and Freshwater Ecosystems. J. Hazard. Mater. 141021. https://doi.org/10.1016/j.jhazmat.2025.141021

Cui, X., Yang, T., Li, Z., & Nowack, B. 2024. Meta-analysis of the hazards of microplastics in freshwaters using species sensitivity distributions. Journal of Hazardous Materials, 463, 132919. https://doi.org/10.1016/j.jhazmat.2023.132919

de Ruijter, V. N., Redondo-Hasselerharm, P. E., Gouin, T., & Koelmans, A. A. 2020. Quality Criteria for Microplastic Effect Studies in the Context of Risk Assessment: A Critical Review. Environmental Science & Technology, 54(19), 11692–11705. https://doi.org/10.1021/acs.est.0c03057

Devipriya, K., Tlija, M., Kumar, C.K., Kumar, V.C., Jana, S., Jana, C., 2025. Automated micro-plastic detection and classification using deep convolution neural network pre-trained models and transfer learning. AIP Adv. 15. https://doi.org/10.1063/5.0243976

Ding, R., Chen, Y., Shi, X., Li, Y., Yu, Y., Sun, Z., Duan, J., 2024. Size-dependent toxicity of polystyrene microplastics on the gastrointestinal tract: Oxidative stress related-DNA

damage and potential carcinogenicity. Sci. Total Environ. 912, 169514. https://doi.org/10.1016/j.scitotenv.2023.169514

Elith, J., Leathwick, J. R., & Hastie, T. 2008. A working guide to boosted regression trees. Journal of Animal Ecology, 77(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

European Chemicals Agency (EU body or agency). 2016. New approach methodologies in regulatory science: Proceedings of a scientific workshop: Helsinki, 19 20 April 2016. Publications Office of the European Union. https://data.europa.eu/doi/10.2823/543644

Fisher, A., Rudin, C., Dominici, F., 2019. 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. 20, 1–88.

Gong, L., Martinez, O., Mesquita, P., Kurtz, K., Xu, Y., Lin, Y., 2023. A microfluidic approach for label-free identification of small-sized microplastics in seawater. Sci. Rep. 13, 1–14. https://doi.org/10.1038/s41598-023-37900-9

Hartmann, N. B., Hüffer, T., Thompson, R. C., Hassellöv, M., Verschoor, A., Daugaard, A. E., Rist, S., Karlsson, T., Brennholt, N., Cole, M., Herrling, M. P., Hess, M. C., Ivleva, N. P., Lusher, A. L., & Wagner, M. 2019. Are We Speaking the Same Language? Recommendations for a Definition and Categorization Framework for Plastic Debris. Environmental Science & Technology, 53(3), 1039–1047. https://doi.org/10.1021/acs.est.8b05297

Hodkovicova, N., Hollerova, A., Svobodova, Z., Faldyna, M., & Faggio, C. 2022. Effects of plastic particles on aquatic invertebrates and fish – A review. Environmental Toxicology and Pharmacology, 96, 104013. https://doi.org/10.1016/j.etap.2022.104013

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

Kennedy, S.B., Antonio Vital, A.L., Kukkola, A., Miller, E., Yeh, A., Coffin, S., Ahmed, T.K., Bertrand, L., Barrick, A., Cowger, W., Mair, M.M., Doyle, D., 2026. Trends in quality and risk assessment applicability of microplastic ecotoxicity studies. J. Hazard. Mater. Adv. 21, 100942. https://doi.org/10.1016/j.hazadv.2025.100942

Khanam, M.M., Uddin, M.K., Kazi, J.U., 2025. Advances in machine learning for the detection and characterization of microplastics in the environment. Front. Environ. Sci. 13, 1–15. https://doi.org/10.3389/fenvs.2025.1573579

Koelmans, A.A., Redondo-Hasselerharm, P.E., Mohamed Nor, N.H., Kooi, M., 2020. Solving the Nonalignment of Methods and Approaches Used in Microplastic Research to

584 Characterize Risk Consistently. Environ. Sci. Technol. 54, 12307–12315.
585 https://doi.org/10.1021/acs.est.0c02982
586 Koelmans, A.A., Gebreyohanes Belay, B.M., Mintenig, S.M., Mohamed Nor, N.H.,
587 Redondo-Hasselerharm, P.E., de Ruijter, V.N., 2023. Towards a rational and efficient risk
588 assessment for microplastics. TrAC - Trends Anal. Chem. 165, 117142.
589 https://doi.org/10.1016/j.trac.2023.117142
590 Koelmans, A.A., Redondo-Hasselerharm, P.E., Nor, N.H.M., de Ruijter, V.N.,
591 Mintenig, S.M., Kooi, M., 2022. Risk assessment of microplastic particles. Nat. Rev. Mater. 7,
592 138–152. https://doi.org/10.1038/s41578-021-00411-y
593 Lambert, S., Scherer, C., & Wagner, M. 2017. Ecotoxicity testing of microplastics:
594 Considering the heterogeneity of physicochemical properties. Integrated Environmental
595 Assessment and Management, 13(3), 470–475. https://doi.org/10.1002/ieam.1901
596 Li, Z., Zhang, Y., Chen, Y., 2025. Toxicity of nanoplastics: machine learning combined
597 with meta-analysis. Nanoscale Horizons 10, 2017–2036. https://doi.org/10.1039/d5nh00024f
598 Liu, C., Zong, C., Chen, S., Chu, J., Yang, Y., Pan, Y., Yuan, B., Zhang, H., 2024.
599 Machine learning-driven QSAR models for predicting the cytotoxicity of five common
600 microplastics. Toxicology 508, 153918. https://doi.org/10.1016/j.tox.2024.153918
601 Mayer, M. 2024. missRanger: Fast Imputation of Missing Values (Version 2.6.1)
602 [Computer software]. https://cran.r-project.org/web/packages/missRanger/index.html
603 Mehinto, A.C., Coffin, S., Koelmans, A.A., Brander, S.M., Wagner, M., Thornton
604 Hampton, L.M., Burton, A.G., Miller, E., Gouin, T., Weisberg, S.B., Rochman, C.M., 2022.
605 Risk-based management framework for microplastics in aquatic ecosystems. Microplastics and
606 Nanoplastics 2. https://doi.org/10.1186/s43591-022-00033-3
607 Menzel, T., Meides, N., Mauel, A., Mansfeld, U., Kretschmer, W., Kuhn, M., Herzig,
608 E. M., Altstädt, V., Strohriegl, P., Senker, J., & Ruckdäschel, H. 2022. Degradation of low-
609 density polyethylene to nanoplastic particles by accelerated weathering. Science of The Total
610 Environment, 826, 154035. https://doi.org/10.1016/j.scitotenv.2022.154035
611 Meyers, N., De Witte, B., Schmidt, N., Herzke, D., Fuda, J.L., Vanavermaete, D.,
612 Janssen, C.R., Everaert, G., 2024. From microplastics to pixels: testing the robustness of two
613 machine learning approaches for automated, Nile red-based marine microplastic identification.
614 Environ. Sci. Pollut. Res. 31, 61860–61875. https://doi.org/10.1007/s11356-024-35289-0
615 Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N.,
616 Patzlaff, H., Naumann, F., Harmouch, H., 2025. The effects of data quality on machine learning
617 performance on tabular data. Inf. Syst. 132, 102549. https://doi.org/10.1016/j.is.2025.102549

Monclús, L., Arp, H.P.H., Groh, K.J., Faltynkova, A., Løseth, M.E., Muncke, J., Wang, Z., Wolf, R., Zimmermann, L., Wagner, M., 2025. Mapping the chemical complexity of plastics. Nature 643, 349–355. https://doi.org/10.1038/s41586-025-09184-8

Müller, S., 2025. How to crack a SMILES: automatic crosschecked chemical structure resolution across multiple services using MoleculeResolver. J. Cheminform. 17. https://doi.org/10.1186/s13321-025-01064-7

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

R Core Team. 2024. _R: A Language and Environment for Statistical Computing_. (Vienna, Austria) [Computer software]. R Foundation for Statistical Computing.

Ramsperger, A. F. R. M., Jasinski, J., Völkl, M., Witzmann, T., Meinhart, M., Jérôme, V., Kretschmer, W. P., Freitag, R., Senker, J., Fery, A., Kress, H., Scheibel, T., & Laforsch, C. 2022. Supposedly identical microplastic particles substantially differ in material properties, influencing particle-cell interactions and cellular responses. Journal of Hazardous Materials, 425, 127961. https://doi.org/10.1016/j.jhazmat.2021.127961

Ramsperger, A. F. R. M., Narayana, V. K. B., Gross, W., Mohanraj, J., Thelakkat, M., Greiner, A., Schmalz, H., Kress, H., & Laforsch, C. 2020. Environmental exposure enhances the internalization of microplastic particles into cells. Science Advances, 6(50), eabd1211. https://doi.org/10.1126/sciadv.abd1211

Renftle, M., Trittenbach, H., Poznic, M., Heil, R., 2024. What do algorithms explain? The issue of the goals and capabilities of Explainable Artificial Intelligence (XAI). Humanit. Soc. Sci. Commun. 11, 1–10. https://doi.org/10.1057/s41599-024-03277-x

Rochman, C.M., Brookson, C., Bikker, J., Djuric, N., Earn, A., Bucci, K., Athey, S., Huntington, A., McIlwraith, H., Munno, K., Frond, H. De, Kolomijeca, A., Erdle, L., Grbic, J., Bayoumi, M., Borrelle, S.B., Wu, T., Santoro, S., Werbowski, L.M., Zhu, X., Giles, R.K., Hamilton, B.M., Thaysen, C., Kaura, A., Klasios, N., Ead, L., Kim, J., Sherlock, C., Ho, A., Hung, C., 2019. Rethinking microplastics as a diverse contaminant suite. Environ. Toxicol. Chem. 38, 703–711. https://doi.org/10.1002/etc.4371

Rudin, C. 2019. Stop explaining black-box machine learning models for high-stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Schrank, I., Trotter, B., Dummert, J., Scholz-Böttcher, B. M., Löder, M. G. J., & Laforsch, C. 2019. Effects of microplastic particles and leaching additive on the life history and

morphology of Daphnia magna. Environmental Pollution, 255, 113233. https://doi.org/10.1016/j.envpol.2019.113233

Schwarzer, M., Brehm, J., Vollmer, M., Jasinski, J., Xu, C., Zainuddin, S., Fröhlich, T., Schott, M., Greiner, A., Scheibel, T., & Laforsch, C. 2022. Shape, size, and polymer-dependent effects of microplastics on Daphnia magna. Journal of Hazardous Materials, 426, 128136. https://doi.org/10.1016/j.jhazmat.2021.128136

Seewoo, B.J., Goodes, L.M., Mofflin, L., Mulders, Y.R., Wong, E.V., Toshniwal, P., Brunner, M., Alex, J., Johnston, B., Elagali, A., Gozt, A., Lyle, G., Choudhury, O., Solomons, T., Symeonides, C., Dunlop, S.A., 2023. The plastic health map: A systematic evidence map of human health studies on plastic-associated chemicals. Environ. Int. 181, 108225. https://doi.org/10.1016/j.envint.2023.108225

Tang, Y., Liu, Y., Chen, Y., Zhang, W., Zhao, J., He, S., Yang, C., Zhang, T., Tang, C., Zhang, C., & Yang, Z. 2021. A review: Research progress on microplastic pollutants in aquatic environments. Science of The Total Environment, 766, 142572. https://doi.org/10.1016/j.scitotenv.2020.142572

Thompson, R. C., Courtene-Jones, W., Boucher, J., Pahl, S., Raubenheimer, K., & Koelmans, A. A. 2024. Twenty years of microplastics pollution research—What have we learned? Science, 0(0), eadl2746. https://doi.org/10.1126/science.adl2746

Thornton Hampton, L. M., Lowman, H., Coffin, S., Darin, E., De Frond, H., Hermabessiere, L., Miller, E., De Ruijter, V. N., Faltynkova, A., Kotar, S., Monclús, L., Siddiqui, S., Völker, J., Brander, S., Koelmans, A. A., Rochman, C. M., Wagner, M., & Mehinto, A. C. 2022. A living tool for the continued exploration of microplastic toxicity. Microplastics and Nanoplastics, 2(1), 13. https://doi.org/10.1186/s43591-022-00032-4

Thornton Hampton, L.M., Wyler, D.B., Almroth, B.C., Coffin, S., Cowger, W., Doyle, D., Hataley, E.K., Hutton, S.J., Mair, M.M., Miller, E.L., Monclús, L., Sharpe, E.E., Samreen, S., Ahmed, K.T., Allamby, Q.P. V, Vital, A.L.A., 2025. The Toxicity of Microplastics Explorer (ToMEx) 2.0. Microplastics and Nanoplastics 5, 12. https://doi.org/10.1186/s43591-025-00145-6

Tran, H.T., Hadi, M., Nguyen, T.T.H., Hoang, H.G., Nguyen, M.K., Nguyen, K.N., Vo, D.V.N., 2023. Machine learning approaches for predicting microplastic pollution in peatland areas. Mar. Pollut. Bull. 194, 115417. https://doi.org/10.1016/j.marpolbul.2023.115417

Tropsha, A., 2010. Best practices for QSAR model development, validation, and exploitation. Mol. Inform. 29, 476–488. https://doi.org/10.1002/minf.201000061

Wang, M., Wang, C., Yang, X., Mu, Y., Zhao, X., Wu, F., Leung, K. M. Y., Shin, H.-M., Sayes, C. M., & Giesy, J. P. 2026. Machine-Learning-Accelerated Prediction of Water Quality Criteria for Microplastics. ACS ES&T Water, acsestwater.5c00602. https://doi.org/10.1021/acsestwater.5c00602

Wieland, S., Ramsperger, A. F. R. M., Gross, W., Lehmann, M., Witzmann, T., Caspari, A., Obst, M., Gekle, S., Auernhammer, G. K., Fery, A., Laforsch, C., & Kress, H. 2024. Nominally identical microplastic models differ greatly in their particle-cell interactions. Nature Communications, 15(1), 922. https://doi.org/10.1038/s41467-024-45281-4

Wieland, S., Balmes, A., Bender, J., Kitzinger, J., Meyer, F., Ramsperger, A.F., Roeder, F., Tengelmann, C., Wimmer, B.H., Laforsch, C., Kress, H., 2022. From properties to toxicity: Comparing microplastics to other airborne microparticles. J. Hazard. Mater. 428, 128151. https://doi.org/10.1016/j.jhazmat.2021.128151

Yu, F., Hu, X., 2022. Machine learning may accelerate the recognition and control of microplastic pollution: Future prospects. J. Hazard. Mater. 432, 128730. https://doi.org/10.1016/j.jhazmat.2022.128730

Yu, Y., Kumar, M., Bolan, S., Padhye, L.P., Bolan, N., Li, S., Wang, L., Hou, D., Li, Y., 2024. Various additive release from microplastics and their toxicity in aquatic environments. Environ. Pollut. 343, 123219. https://doi.org/10.1016/j.envpol.2023.123219

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

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

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

Table S1. Number of data points for the training, validation, and test subsets of the four datasets 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

Table S2. Number of particle identifiers (particle ID) and data points for the training, validation, 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

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

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

	Imputed	Complete-case	High-quality	Low-quality
Training (5-fold CV average)	AUC = 0.867	AUC = 0.862	AUC = 0.866	AUC = 0.831
Validation	AUC = 0.871	AUC = 0.868	AUC = 0.885	AUC = 0.892
Test	AUC = 0.896	AUC = 0.868	AUC = 0.929	AUC = 0.838

Table S5. Area under the curve (AUC) values for training, validation, and test sets across the 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 = 0.735	AUC = 0.736	AUC = 0.740	AUC = 0.668
Validation	AUC = 0.703	AUC = 0.736	AUC = 0.679	AUC = 0.453
Test	AUC = 0.689	AUC = 0.672	AUC = 0.870	AUC = 0.747

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

Cross-validation	Dataset	Threshold	Sensitivity	MCC
Random	Imputed	0.268	0.850	0.595
	Complete-case	0.275	0.890	0.573
	High-quality	0.290	0.932	0.690
	Low-quality	0.560	0.653	0.570
Blocked	Imputed	0.253	0.798	0.252
	Complete-case	0.481	0.516	0.289
	High-quality	0.207	0.924	0.618
	Low-quality	0.425	0.784	0.398

Table S7. Feature-group ablation analysis under the random cross-validation scheme with high-quality dataset.

Categories	Mean ± standard deviation
Concentration_Endpoints_Experiment_Microplastic_Species	0.908 ± 0.025
Concentration_Endpoints_Experiment_Microplastic	0.907 ± 0.019
Concentration_Endpoints_Microplastic_Species	0.901 ± 0.014
Concentration_Endpoints_Microplastic	0.898 ± 0.020
Concentration_Endpoints_Experiment_Species	0.896 ± 0.022
Concentration_Endpoints_Experiment	0.895 ± 0.021
Concentration_Endpoints_Species	0.891 ± 0.021
Endpoints_Experiment_Microplastic_Species	0.873 ± 0.025
Endpoints_Microplastic_Species	0.869 ± 0.023
Endpoints_Experiment_Microplastic	0.869 ± 0.028
Endpoints_Microplastic	0.866 ± 0.022
Concentration_Endpoints	0.865 ± 0.033

Endpoints_Experiment	0.857 ± 0.036
Endpoints_Experiment_Species	0.849 ± 0.032
Endpoints_Species	0.848 ± 0.013
Concentration_Experiment_Microplastic	0.845 ± 0.023
Concentration_Experiment_Microplastic_Species	0.844 ± 0.029
Concentration_Experiment_Species	0.842 ± 0.028
Concentration_Experiment	0.840 ± 0.030
Concentration_Microplastic_Species	0.833 ± 0.032
Concentration_Species	0.822 ± 0.042
Endpoints	0.815 ± 0.033
Experiment_Microplastic_Species	0.814 ± 0.032
Concentration_Microplastic	0.813 ± 0.032
Experiment_Microplastic	0.810 ± 0.023
Microplastic_Species	0.796 ± 0.027
Experiment	0.793 ± 0.031
Experiment_Species	0.793 ± 0.033
Species	0.772 ± 0.031
Concentration	0.760 ± 0.030
Microplastic	0.759 ± 0.043

Table S8. Feature-group ablation analysis under the blocked cross-validation scheme with high-quality dataset.

Categories	**Mean ± standard deviation**
Concentration_Endpoints_Experiment_Species	0.782 ± 0.098
Concentration_Endpoints_Experiment	0.779 ± 0.063
Concentration_Endpoints_Experiment_Microplastic	0.776 ± 0.096
Endpoints_Species	0.758 ± 0.086
Concentration_Endpoints_Experiment_Microplastic_Species	0.750 ± 0.083
Concentration_Experiment	0.750 ± 0.058
Concentration_Endpoints_Species	0.746 ± 0.098
Concentration_Endpoints_Microplastic_Species	0.739 ± 0.105
Endpoints_Experiment_Microplastic_Species	0.738 ± 0.103
Concentration_Experiment_Microplastic_Species	0.732 ± 0.060

Endpoints	0.727 ± 0.069
Concentration_Microplastic_Species	0.716 ± 0.091
Endpoints_Experiment_Species	0.715 ± 0.124
Experiment_Microplastic	0.708 ± 0.134
Concentration_Endpoints	0.706 ± 0.095
Concentration_Endpoints_Microplastic	0.706 ± 0.139
Endpoints_Microplastic_Species	0.706 ± 0.082
Endpoints_Experiment	0.705 ± 0.098
Endpoints_Experiment_Microplastic	0.703 ± 0.100
Concentration_Species	0.703 ± 0.091
Concentration_Experiment_Microplastic	0.701 ± 0.072
Concentration_Experiment_Species	0.701 ± 0.146
Species	0.698 ± 0.101
Endpoints_Microplastic	0.671 ± 0.094
Experiment_Microplastic_Species	0.670 ± 0.077
Microplastic_Species	0.658 ± 0.098
Experiment	0.638 ± 0.106
Experiment_Species	0.632 ± 0.121
Microplastic	0.604 ± 0.120
Concentration_Microplastic	0.571 ± 0.088
Concentration	0.566 ± 0.084