QSPR and q-RASPR predictions of the adsorption capacity of polyethylene, polypropylene and polystyrene microplastics for various organic pollutants in diverse aqueous environments

Md Mobarak Hossain[1], Arkaprava Banerjee[1], Mainak Chatterjee[1], Kunal Roy[1, *], and Mark TD Cronin[2]

[1]Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India

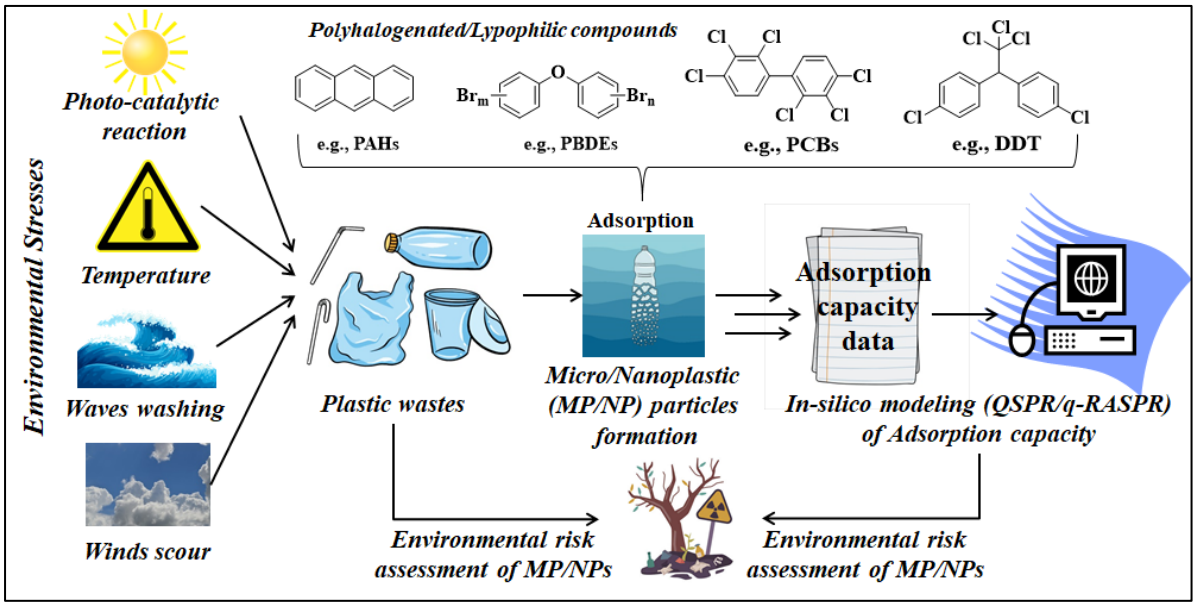
[2]School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Liverpool L3 3AF, UK

***Corresponding author**
Email: kunal.roy@jadavpuruniversity.in

ABSTRACT

Microplastics (MPs) and nanoplastics (NPs) have become significant environmental concerns due to their widespread presence in various ecosystems and their potential to act as carriers and accumulators of organic pollutants, including pesticides, pharmaceuticals and industrial chemicals, many of which are known to cause harm to aquatic organisms and bioaccumulate in the food chain. The adsorption of organic pollutants onto microplastics is a complex process influenced by various factors, including the physicochemical properties of both the microplastics and the pollutants. Understanding this process is crucial for assessing the environmental impact of microplastics and developing effective mitigation strategies. In this study, experimental data on the adsorption of organic pollutants onto microplastics in different aqueous environments were used to develop quantitative structure-property relationship (QSPR) and quantitative read-across structure-property relationship (q-RASPR) models establishing quantitative relationships between the chemical structures of the pollutants and their adsorption behavior, providing insights into the key molecular properties that govern the adsorption process. Apart from developing models separately for different categories of microplastics in a particular aqueous environment, global models were also developed combining all the data points. External predictions from q-RASPR models were more accurate than those from QSPR models, and the final global q-RASPR model, with only one descriptor, showed excellent statistical fit. These models will be helpful for planning pollution control measures and the sustainable management of plastic waste and environmental contamination.

Keywords: Microplastics, Nanoplastics, Organic pollutants, Adsorption capacity, Environmental risk, q-RASPR.



Graphical abstract

Environmental significance

Microplastics (MPs) and nanoplastics (NPs) act as carriers and accumulators of organic pollutants, including pesticides, pharmaceuticals and industrial chemicals, many of which are known to cause harm to aquatic organisms and bioaccumulate in the food chain. The adsorption of organic pollutants onto microplastics is a complex process influenced by various factors. In this study, experimental data on the adsorption of organic pollutants onto microplastics in different aqueous environments were used to develop quantitative structure-property relationship (QSPR) and quantitative read-across structure-property relationship (q-RASPR) models establishing quantitative relationships between the chemical structures of the pollutants and their adsorption behavior, providing insights into the key molecular properties that govern the adsorption process.

1. INTRODUCTION

The increasing production of plastics is estimated to reach 250 million tons by 2025 (1). It was estimated that in 2016 alone, up to 23 million tons of plastics entered into the aquatic ecosystems (2). This amount will double by 2030 (2) and is projected to increase by 300–400% by 2050 (3). Studies have shown that 1.5–4.5% of globally produced plastics were released directly into the ocean (4,5). The plastics remain undegraded in the environment for a long time. When discarded, plastic residues accumulate in the environment, and due to environmental stresses (e.g., photonic radiation, temperature, wave wash, air pressure, etc.), they gradually break down into tiny particles of plastic smaller than 5 mm in size. These particles are known as microplastics (MPs). (7). Nanoplastics (NPs) are tinier than MPs and their size range vary from 1 to 1000 nm. The primary source of MPs is industry-originated micro-sized plastic particles; these particles and other plastic waste also get fragmented in the environment to form MPs (8). MPs, which do not mix with water, can be converted to NPs upon the addition of certain chemical additives and vigorous mixing; therefore, the NPs found in the environment are mostly engineered products discarded after use. The small particle size of MPs (less than 5 mm) and NPs (1-1000 nm) allows easy transportation by wind and water, leading to their widespread distribution across terrestrial and aquatic environments. As a result, MPs and NPs have been found in diverse ecosystems, including oceans, rivers, lakes, soil, and even in the atmosphere and may cause physical and toxicological damage to the environmental organisms and humans (6).

The environmental impact of these minute plastic particles is a growing area of concern, as MPs and NPs can be ingested to a wide range of organisms, from zooplankton to larger marine animals. This ingestion can lead to physical harm, hindered feeding ability and potential transfer of toxic substances to higher levels of the food chain (biomagnification). Furthermore, MPs/NPs can act as carriers for pollutants and other harmful substances, posing risks to wildlife and human health. (9) Humans are also directly in contact with MPs/NPs through inhalation, consumption of food/water, and even by skin contact (e.g. personal care products and mobile phone cases), and they are exposed to more plastics than ever before. Therefore, the amount of MPs/NPs residues in the human body is increasing as environmental plastic pollution worsens. MPs and NPs have been detected in many body tissues including the human placenta and the eye. (10) (11) The placenta may be highly exposed to MPs/NPs due to its high blood flow and

absorption of nutrients from the mother's blood and through the placental barrier. (12) Because of the presence of MP/NPs in the placenta, possible health impact on fetal development is also a serious concern. (13) In another study, it has been observed that participants who had microplastics in their blood arteries had an increased risk of heart attack, stroke, or early death compared to those who had not. (14)

Due to the comparatively large surface area of MPs and NPs, they adsorb different organic pollutants and other contaminants from the aquatic environment leading to an increased risk of potential hazards. Thus, to understand the toxic potential of MPs/NPs, a proper assessment of their adsorption capacity is necessary. Generally, the adsorption capacity of MPs/NPs is represented by the equilibrium partitioning coefficient (K_d) of a particular substance between MPs/NPs and water. (15) K_d is calculated as:

$$K_d = q_e/c_e \quad (1)$$

where q_e and c_e represent the concentration of adsorbed substances in microplastics and water respectively. (16-17)

The K_d (kg/L) value provides information about how effectively MPs/NPs can adsorb organic pollutants from the surrounding water. (18) K_d is influenced by various factors, including the characteristics of the aqueous environment and the properties of the microplastics themselves. To be more precise, factors such as pH, temperature, salinity, and the presence of other substances in the water can affect the K_d value. Additionally, the chemical composition, size, and shape of the MPs/ NPs can also impact their adsorption capacity. Among these factors, the chemical nature of MPs/ NPs and the aqueous environment are the most important variables that should be considered when measuring K_d values. (19) The measurement of adsorption capacity is performed using adsorption equilibrium equipment, but the additional variables like temperature, pressure, surface area, pore size distribution, pH, solution composition, contact time, etc. make this process more complex and time-consuming. Currently, research on the adsorption capacity of MPs/ NPs is limited, and there is a lack of comprehensive data regarding their adsorption behavior. This scarcity of data limits our ability to fully understand and estimate the risks associated with micro/nanoplastics in the environment. As a result, there is a need for a rapid and precise method to determine K_d values under various adsorption situations. Developing such methods would greatly aid researchers in understanding the behavior of MPs/NPs and their interactions with organic pollutants in different environmental conditions. This would ultimately contribute to better risk assessment and management of MPs/ NPs in the future. (18)

In silico studies, which involve computer-based simulations and modeling of physical and biological systems, offer several advantages over experimental studies in certain contexts. (20) Firstly, they are cost-effective, requiring fewer resources and materials, making them beneficial for large-scale or high-throughput analyses. Secondly, *in silico* studies are time-efficient, allowing rapid testing and iteration of hypotheses, which is advantageous in time-critical fields such as drug discovery and regulatory toxicology. Moreover, many *in silico* studies are based on publicly available data, making the analyses easier to conduct without extensive laboratory work. They also aid in hypothesis generation by simulating complex biological processes and predicting potential outcomes, guiding experimental research. (18) Furthermore, *in silico* predictions obtained from tools such as Quantitative Structure-Activity/Property Relationships (QSAR/QSPR) and read-across (RA) may be applied in regulatory decision-making by several

regulatory bodies such as the United States Environmental Protection Agency (US EPA), European Chemicals Agency (ECHA), and the European Food Safety Authority (EFSA), etc. (20)

In traditional QSAR/QSPR modeling, one or more attributes (dependent variables) are predicted based on one or more descriptors or features (independent variables) that capture the structural and physicochemical characteristics of the compounds. QSPR models are trained using experimental data to establish relationships between the chemical structure and the properties of interest. (21) On the other hand, RA is a similarity-based method that attempts to interpolate a compound's effects or properties by identifying one or more similar compound(s) with adequate data. (22-23) For every query (or target) compound in RA, similar compound(s) are identified. RA can be based on chemical categories (i.e. two or more source molecules) or the analogue approach (one target molecule). (24) The Read-Across Structure-Property Relationship (RASPR) method is an advanced predictive modeling approach that combines the principles of both supervised learning, specifically QSPR, and unsupervised learning, particularly RA. (25) The goal of RASPR is to create a more accurate and reliable model for predicting the properties of chemicals, such as their activity, property, or toxicity. (26) The key concept behind RASPR is to couple significant structural and physicochemical characteristics with error-based measurements and similarity information obtained from RA. (26) The quantitative RASPR (q-RASPR) method specifically emphasizes the use of quantitative measurements and metrics to integrate supervised and unsupervised learning components. By incorporating both experimental data-driven QSPR modeling and similarity-based RA, q-RASPR seeks to enhance the accuracy and reliability of property predictions for chemical compounds.

In this current investigation, we have used K_d (kg/L) values (27) for the three most commonly found MPs, namely polyethylene (PE), polypropylene (PP), and polystyrene (PS). K_d values were reported in diverse aqueous environments and modeled to create predictive approaches for the adsorption capacity of MPs. The main objectives of this study were (1) to develop local QSPR/q-RASPR models for the prediction of the adsorption capacity of MPs (specific for a microplastic and a particular aqueous environment), (2) to develop global QSPR/q-RASPR models by incorporating an appropriate dummy variable for the effective predictions of adsorption capacity irrespective of the chemical nature of MPs and the aqueous environment, (3) to analyse the QSPR and q-RASPR statistically in the light of the quality of predictions, and (4) to identify the chemical features of organic pollutants that influence the adsorption on MP/NPs.

2. MATERIAL AND METHODS

2.1. Data collection and curation

The experimental data sets refer to those where the adsorption capacity (Kd values in kg/L) of microplastics (MPs) for organic contaminants was directly measured through physical experiments using various methods. The adsorption capacity (K_d in kg/L) values can be measured experimentally using various methods, depending on the specific system and conditions being studied. Some common methods for measuring K_d values include: Batch equilibrium method, Column chromatography method, Solid phase extraction (SPE) method, Surface plasmon resonance (SPR) method, etc. In this study, the adsorption capacity (K_d in kg/L) data of three MPs were collected from the literature. (28) No theoretical or predicted data has been used for model development in this study. The collected data were organised into five

local datasets and three global datasets based on the type of MPs and the aqueous environment. In the first three local datasets (Dataset 1, 2, and 3), experimental K_d values of 34, 24, and 47 organic contaminants on PE MPs in seawater, freshwater, and pure water environment, respectively, were gathered. In datasets 4 and 5, experimental K_d values of 32 and 25 organic contaminants on PP and PS MPs respectively (both in a seawater environment), were collected. There is no mixing of two variables (the chemical nature of MPs and the aqueous environment) in the data points of the five local datasets. In the case of data sets 1, 4, and 5, three isomeric compounds were removed because those compounds could not be distinguished by 2D descriptors. In Global Set 1(for sea water with different type of microplastics), data points of Datasets 1, 4, and 5 have been amalgamated to introduce variability in the type of MPs. Likewise, in Global Set 2 (polyethylene microplastics in different types of water), data points of Datasets 1, 2, and 3 have been amalgamated to incorporate variability in the aqueous environment. In the final dataset (Global Set 3 for different microplastics in different water environments), we have amalgamated all the local datasets (Datasets 1-5) to introduce variability in the chemical nature of MPs as well as the aqueous environment. Global Sets 1, 2, and 3 contain 91, 105, and 162 data points respectively. The data distribution among these 8 datasets has been graphically presented in **Figure 1**. Due to the amalgamation of data points into Global sets, there are some repetitions of organic pollutants, these are differentiated using indicator variables for the MPs or the environment. A summary of the tabulated data is provided in **Table 1**. All datasets, along with their observed adsorption capacity data and details of indicator variables are tabulated in **Supplementary Materials 1**. The data comprises several classes of organic pollutants including polychlorinated biphenyls, polycyclic aromatic hydrocarbons, aromatic hydrocarbons, chlorobenzenes, hexachlorocyclohexanes, aliphatic hydrocarbons, antibiotics, and perfluorinated compounds. *Log* K_d (Kg/L) values of all data points have been used as endpoints of the subsequent modeling in this study to minimize the numeric range of adsorption capacity values. In the Global dataset 3, we have found 6 compounds which are getting adsorbed onto PE, PP, and PS microplastics. We have plotted their adsorption capacity (*log* K_d) in the form of a bar diagram (**Figure S1** of **Supplementary Material 2**); in general, the adsorption capacity of polypropylene microplastics is lower than that of polyethylene and postyrene microplastics for the selected compounds.

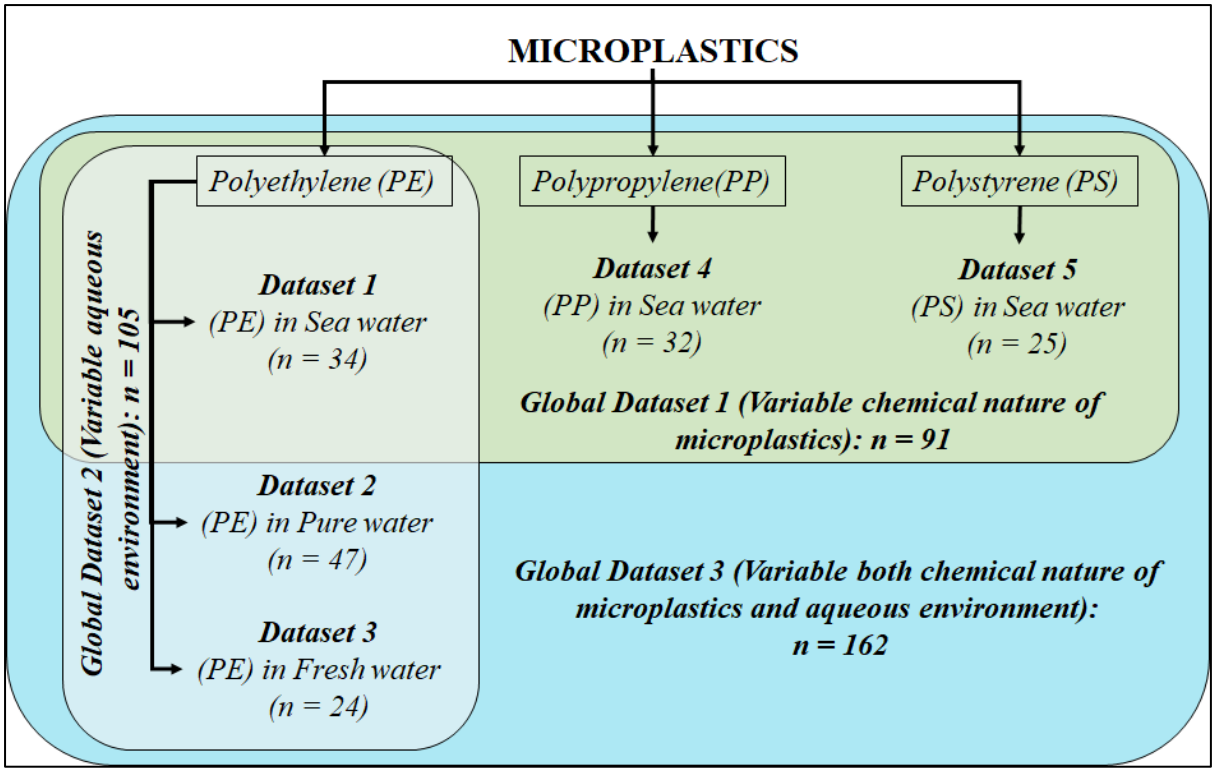


Figure 1 The data sets of the adsorption capacity of microplastics in diverse water environments

2.2. Molecular descriptor calculation and data pre-treatment

As defined by Todeschini and Consonni, a descriptor is "*the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiments*". (29) In simpler words, the quantitative values obtained from the structural details of molecules are called molecular descriptors. In this study, alvaDesc v2.0.6 was used to calculate various classes of 2D descriptors, including molecular properties, 2D atom pairs, atom type E-state indices, atom-centered fragments, functional group counts, connectivity indices, ring descriptors, constitutional indices, and extended topochemical atom (ETA) indices. (30) As demonstrated in several previous QSPR and q-RASPR studies, these classes of descriptors are highly interpretable and effective in the model-building process. (31), (32), (33) The computed descriptors were then screened through a pre-treatment procedure using the Java-based tool DataPreTreatmentGUI 1.2, which can be downloaded from https://teqip.jdvu.ac.in/QSAR_Tools/. The goal was to eliminate the redundant and inter-correlated descriptors with a correlation coefficient cut-off value of 0.95 and a variance cut-off of 0.001. Descriptors with constant or null values for every data point as well as those with strong inter-correlations among them are eliminated during this phase. **Table 1** shows how many descriptors are present in each data set.

2.3. Dataset splitting

Before developing the model, it is essential to divide the dataset into training and test sets for internal and external validation tests of the developed models. Dataset DivisionGUI1.2, a Java-based tool provided at https://teqip.jdvu.ac.in/QSAR_Tools/, was used to divide the dataset in this work in a ratio of 75: 25. Sorted property-based division, Euclidean distance-based

division, and Kennard-Stone division techniques were used. Following the division, pre-treatment was applied to the training and test sets to eliminate the inter-correlated and constant/near-constant descriptors generated after division using the dataPreTreatmentTrainTest1.0 tool available from https://teqip.jdvu.ac.in/QSAR_Tools/. The training set was used to build a model, while the test set was used to assess the model's prediction performance and obtain external validation. **Table 1** shows the number of compounds in the training and test sets along with the respective division algorithms for the best model selection. (34)

2.4 Feature selection and preliminary QSPR development

In order to identify the most important descriptors for modeling the property, stepwise selection was carried out in this study. (35-37) In general practice, forward stepwise selection starts with the descriptor that has the strongest correlation with the property of interest that is added based on statistical criteria such as *p*-value or *F*-statistic which serves as the fitness function (34-35). The *F-to-enter* value of 4 and the *F-to-remove* value of 3.9 were used as the stepping criteria in this study (38-39) which corresponds to a *t*-value greater than 2 justifying statistical significance. Descriptors should continue to be added or removed one at a time, with the model being re-evaluated at each step to determine if the inclusion or exclusion of descriptors improves the model's performance based on statistical tests and criteria. The addition of descriptors should be stopped when no more descriptors meet the predefined criteria for inclusion in the model. Similarly, descriptors should be removed from the model if they no longer meet the criteria. Once the stepwise selection process is complete, the best combination of descriptors is evaluated using the statistical metrics such as determination coefficient (R^2), adjusted R^2, cross-validated correlation coefficient (Q^2_{CV}), mean absolute error (MAE), and root mean square error (RMSE). (40) After obtaining the best descriptor combination, partial least squares (PLS) regression (41) was applied using the selected descriptors with an optimum number of latent variables (LVs). The number of LVs was optimized based on the X-variance that has been extracted into the LVs from the descriptor space and the value of Q^2_{LOO}.

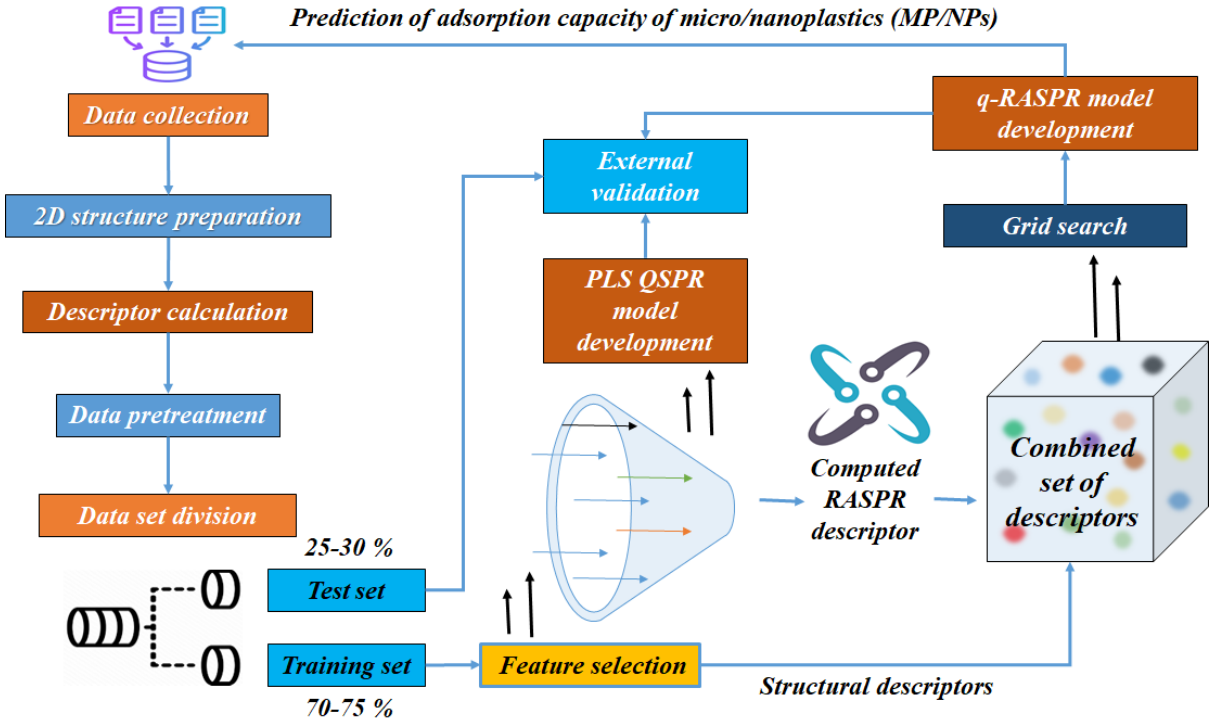


Figure 2 Schematic representation of the complete workflow for QSPR and q-RASPR model development

Table 1. Details of data sets of adsorption capacity data for microplastics and nanoparticles, data set division and the division algorithms

Data set	Type of microplastics	Aqueous Environment	No. of Indicator variables	No. of compounds	No. of calculated descriptors	No. of Training set data points	No. of Test set data points	Division Algorithm
Data set 1	PE	Sea water	0	34	165	26	8	Euclidean Distance-based division
Data set 2	PE	Pure water	0	47	196	36	11	Euclidean Distance-based division
Data set 3	PE	Freshwater	0	24	95	18	06	Sorted-Response-based division
Data set 4	PP	Sea water	0	32	104	24	08	Euclidean Distance-based division
Data set 5	PS	Sea water	0	25	109	19	06	Euclidean Distance-based division
Global set 1	PE+PP+PS	Sea water	3	91	219	69	22	Sorted-Response-based division
Global set 2	PE	Sea water +pure water + fresh water	3	105	320	73	32	Kennard Stone method
Global set 3 (whole data set)	PE+PP+PS	Sea water +pure water + fresh water	6	162	342	122	40	Euclidean Distance-based division

*PE = polyethylene, PP = polypropylene, PS = polystyrene

2.5. Optimization of RA hyperparameters and subsequent RASPR descriptor calculation

RA prediction and associated hyperparameter optimization is the primary step of the q-RASPR modeling algorithm. In this study, RA predictions were performed following the method reported by Chatterjee et al. (42) This method computes the RA predictions for the query compound(s) using three different similarity-based methods: Euclidean distance (ED)-based similarity, Gaussian kernel (GK) function similarity, and Laplacian kernel (LK) function similarity. (43), (44) There are three major hyperparameters (σ, γ, number of close source/training compounds), that should be optimized before the RA predictions. The optimum number of close source compounds is required for the optimum predictions from all these three methods, while σ and γ should be optimized to improve the prediction from Gaussian kernel function-based RA and Laplacian kernel function-based RA respectively. The RA-based hyperparameter optimization was performed using a Java-based tool Auto_RA_Optimizer-v1.0 (available from https://sites.google.com/jadavpuruniversity.in/dtc-lab-software/home) based on only the training set chemical information for all the datasets. The optimized settings of the hyperparameters for all the datasets are shown in the comparison table in **Supplementary Material 1**. These hyperparameters were used to compute the RASPR descriptors for the training and test sets using the tool RASAR-Desc-Calc-v3.0.2. For the calculation of the test set RASPR descriptors, the test sets were used as the query set input files whereas for calculating the training set RASPR descriptors, the training sets were used as the query set input files. In both cases, the training set file served as the source/reference set One point should be noted here that during the computation of training set RASPR descriptors, there is a chance of overfitting and bias due to the use of training set twice as the source/reference and query set input files. To avoid this possibility, the Leave-Same-Out algorithm has been used to compute the training set RASPR descriptors. The 'Leave-Same-Out' algorithm identifies the identical source compounds (based on the X- and Y-variable-based similarities) present in the training set against each query compound and then removes them prior to the RASPR descriptor computation. (45)

2.6. Development of q-RASPR models

After the computation of the RASAR descriptors, the set of 18 RASPR descriptors was pooled with the previously selected structural and physicochemical descriptors to generate complete descriptor pools (Data fusion). These pools were subjected to feature selection using a grid search algorithm that generates different combinations of descriptors and their corresponding Multiple linear regression (MLR) model statistics. Grid search was performed once using one single (best) descriptor and again using the same number of descriptors as that of the corresponding PLS QSPR models. This grid search was carried out using the Best Subset Selection v2.1 tool (46), (47), which is accessible from https://teqip.jdvu.ac.in/QSAR_Tools/. A suitable descriptor combination was selected based on primarily their cross-validation statistics, and the corresponding PLS models were developed. This entire model development procedure was performed for all five individual datasets and the three global datasets. The schematic representation of QSPR and q-RASPR model development is shown in **Figure 2**.

2.7. Statistical validation metrics

A model's validation is crucial to prove its usefulness and choosing whether to use it further. To evaluate the model's quality, goodness-of-fit, robustness, reliability and predictivity, a variety of statistical measures are available. The determination coefficients (R^2 and $R^2_{(adj)}$) are used to

evaluate the statistical fit of the model. Model validation metrics can be divided into two categories: metrics for internal validation and metrics for external validation. To assess a model's robustness and goodness-of-fit, internal validation is limited to the training set; external validation, on the other hand, is conducted on the test set to assess a model's external predictivity. (48) The internal validation metric Q^2_{LOO} (leave-one-out cross-validated correlation coefficient), when applied to the training set, measures a model's robustness. A test set and a training set were created from the original dataset, and the test set was utilized to determine the predictivity of a model by the computation of external validation metrics such as Q^2_{F1}, Q^2_{F2}, and MAE_{test}. (49)

$$Q^2_{F1} = 1 - \frac{\sum(Y_{obs(test)} - Y_{pred(test)})^2}{\sum(Y_{obs(test)} - \bar{Y}_{training})^2} \quad (2)$$

$$Q^2_{F2} = 1 - \frac{\sum(Y_{obs(test)} - Y_{pred(test)})^2}{\sum(Y_{obs(test)} - \bar{Y}_{test})^2} \quad (3)$$

In the Eqs. (2) and (3), $Y_{obs(test)}$ and $Y_{pred(test)}$ are the observed and predicted values of the test data points while $\bar{Y}$ is the mean response value of training or test set compounds.

2.8. Applicability domain analysis (AD) and Y-randomization

A defined domain of applicability indicates the chemical space associated with the QSPR and q-RASPR models and relates to OECD principle 3. (50) The chemicals employed in the model's development (the training set) establish the response and chemical structure space that AD represents. Since error-based metrics do not assess the model's performance considering the mean response, they were utilized to reflect a genuine indication of prediction quality in terms of prediction errors. The leverage approach is a method commonly used for applicability domain analysis in QSAR and q-RASPR modeling (49). Leverage refers to the influence that a data point has on the model, and compounds with high leverage values can significantly impact the model's predictions. In the context of applicability domain analysis, the leverage approach involves calculating the leverage values for each compound in the training set based on their structural features. Compounds with high leverage values are considered to have a greater impact on the QSAR and q-RASPR model and may be more influential in determining its predictive performance. By identifying compounds with high leverage values, we can assess whether these compounds are representative of the chemical space covered by the training set and whether they are outliers that could potentially affect the model's generalizability to new compounds. Compounds with high leverage values that are also located in regions of chemical space not well-represented by the training set may indicate limitations in the model's applicability domain. In addition to calculating leverage values, we can also use leverage plots to visualize the distribution of leverage values across the training set. This can help identify compounds that stand out in terms of their influence on the model and guide decisions on whether to include or exclude certain compounds from the analysis.

Using SIMCA-P software, the Model Y-randomization test was run to determine whether or not the models were generated by chance. Once the applicability domain of the QSAR model has been established, it is important to validate the model using statistical tests such as Y-randomization. In Y-randomization, the response values of the compounds in the training set are randomly shuffled, while keeping their structural information unchanged. The QSAR

model is then rebuilt using the shuffled data, and its performance is compared to the original model. If the performance of the Y-randomized model is significantly worse than the original model, it indicates that the QSAR model is not solely based on chance correlations and has predictive power. This helps to ensure that the QSAR model is robust and not overfitting the training data. (51)

3. RESULTS AND DISCUSSION

3.1. Developed QSPR models and the result of validation

Eight PLS QSPR models were developed to predict the logarithm of the microplastic/water partition coefficients (log K_d) of organic compounds between PE/seawater, PE/pure water, PE/freshwater, PP/seawater, PS/seawater, PE+PP+PS/seawater, PE/ Seawater+ pure water + freshwater, and PE+PP+PS /Sea water +pure water + fresh water. The first five models were local models developed for the prediction of the adsorption capacity of MPs in a more localized environment, while the final three models were global models developed for the prediction of the adsorption capacity of MPs in comparatively more diverse situations considering different chemical nature of MPs and the aqueous environment. Indicator variables were used to describe the type of microplastics and the aqueous environment. All eight models have been showcased in **Box 1** and the quality metrics of the QSPR models have been shown in **Table 2**. All the modelled descriptors have been defined in Table S1 of **Supplementary Materials 2**. All the models are in an acceptable range of validation metrics (R^2 >0.6, Q^2_{LOO} > 0.5, Q^2_{F1}, Q^2_{F2} > 0.5, MAE_{Train} and MAE_{Test} are on the lower side) as per the OECD protocols. We also run Y-randomization test to determine whether or not the models were generated by chance. Y-randomization plot has been presented in **Supplementary Materials 2**. The obtained intercept values of PLS-QSPR global model 1 (R^2 =0.000127 and Q^2 = -0.554), global model 2 (R^2 =0.0113 and Q^2 = -0.58) and global model 3 (R^2 = -0.0165 and Q^2 = -0.252) inferred that the models were not developed by chance. The observed vs. predicted *log* K_d scatter plots of all QSPR models have been provided in the following **Figure 3**. Uniform scattering of training and test data points have been observed across the trend lines which signifies the goodness-of-fit of the models.

BOX 1 Developed PLS QSPR models

Box 1	
Data set 1 (PE in seawater)	$Logk_d = -3.23063 + 0.10577 \times C - 0.11215 \times F05[C-Cl] + 0.59784 \times F08[Cl-Cl] + 0.11497 \times MLOGP2$
Data set 2 (PE in Pure water)	$Logk_d = 1.93649 - 0.09043 \times C - 005 + 0.07577 \times H - 046 - 0.47812 \times NsssCH - 0.49104 \times B05[C-Cl] + 1.09515 \times F03[N-O] + 0.12008 \times MLOGP2$
Data set 3 (PE in fresh water)	$Logk_{\ d} = 0.62087 + 0.45114 \times F07[C-Cl] + 0.56185 \times MLOGP$
Data set 4 (PP in seawater)	$Logk_d = 85.12387 - 76.68738 \times Mi + 0.50925 \times X1v + 0.38651 \times B05[C-Cl] - 0.16438 \times F06[Cl-Cl]$
Data set 5 (PS in seawater)	$Logk_d = -2.89031 + 0.10155 \times C\% - 1.12911 \times B10[C-O] + 0.60389 \times MLOGP$
Global set 1 (PE+PP+PS in seawater)	$Logk_d = -4.28076 + 0.09437 \times C\% - 1.34637 \times B10[C-F] + 0.03992 \times F07[C-C] + 0.3455 \times F08[Cl-Cl] - 0.76006 \times BLTD48 - 0.05799 \times ind1(pe) - 0.96877 \times ind2(pp)$
Global set 2 (PE in Sea water +pure water + fresh water)	$Logk_d = 3.10302 - 0.23182 \times nHet - 0.10482 \times H - 047 + 0.17136 \times SaaaC + 1.21342 \times B03[N-O] + 0.4744 \times F07[O-F] + 0.52764 \times F08[Cl-Cl] + 0.10979 \times MLOGP2$
Global set 3 (Whole data set) (PE+PP+PS in Sea water +pure water + fresh water)	$Logk_d = 2.7838 - 0.43467 \times ind2(pp) - 0.10751 \times nHet + 0.13027 \times SaaaC + -0.22976 \times F07[C-N] + 0.55486 \times F08[Cl-Cl] + 0.08446 \times MLOGP2$

Table 2. Statistical and validation metrics of the PLS QSPR models

Dataset	Latent Variables (LVs) for *PLS* QSPR models	nDesc	R^2_{Train}	Q^2_{LOO}	MAE_{Train}	Q^2_{F1}	Q^2_{F2}	MAE_{Test}
Data set 1	3	4	0.979	0.968	0.257	0.953	0.952	0.303
Data set 2	5	6	0.953	0.912	0.214	0.965	0.964	0.215
Data set3	1	2	0.983	0.979	0.188	0.978	0.977	0.276
Data set 4	3	4	0.967	0.941	0.165	0.939	0.939	0.353
Data set 5	1	3	0.949	0.924	0.331	0.975	0.973	0.308
Global set 1	4	7	0.967	0.955	0.302	0.922	0.921	0.425
Global set 2	5	7	0.951	0.940	0.385	0.901	0.839	0.372
Global set 3 (whole data set)	4	6	0.925	0.915	0.426	0.931	0.930	0.427

nDesc = the number of descriptors

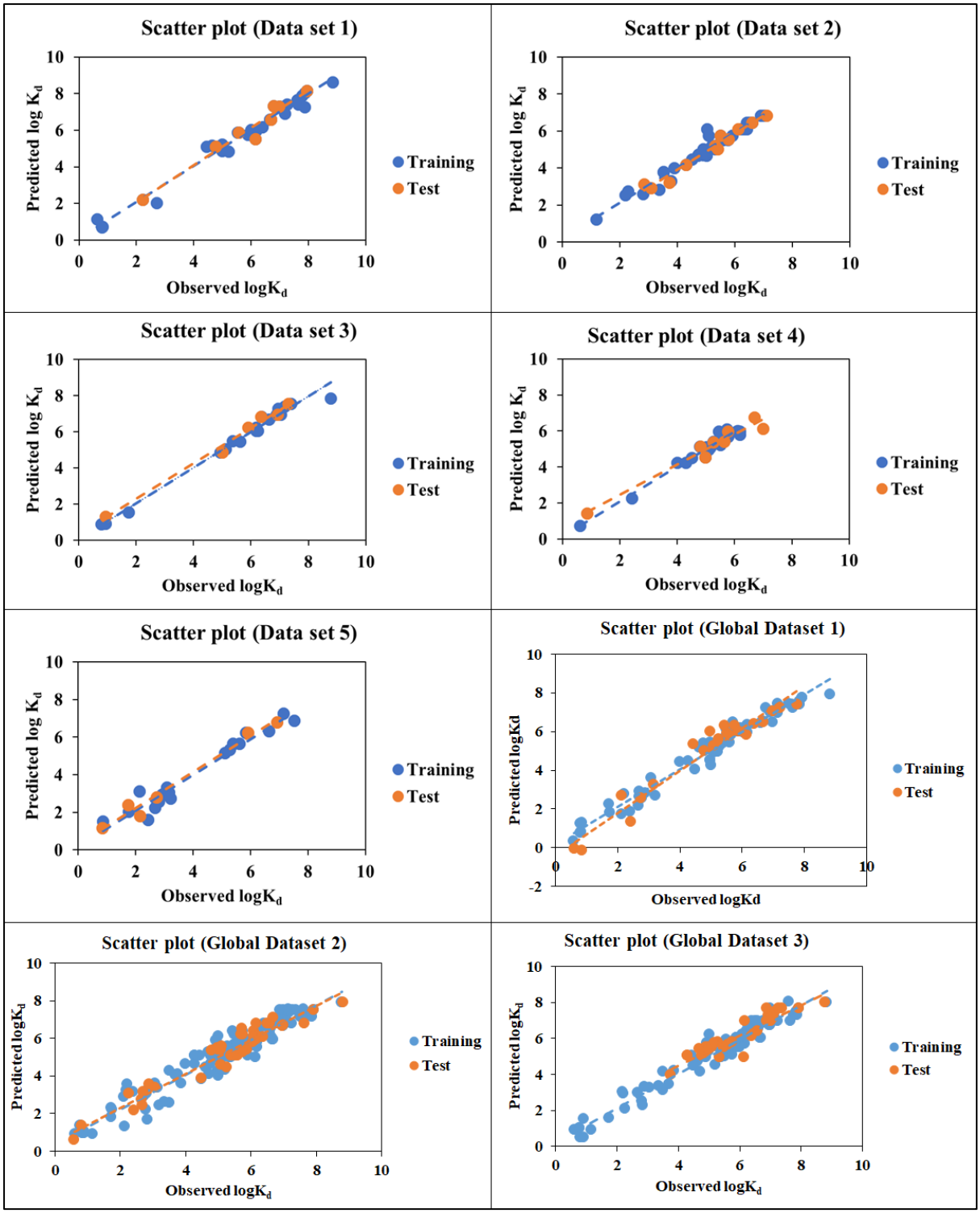


Figure 3 Observed **vs.** Predicted scatter plots of the developed QSPR models (Datasets 1-5, Global Datasets 1-3)

3.2 Developed q-RASPR models and their validation results

The q-RASPR models were developed to check the enhancement in the external predictivity as compared to the QSPR models. In this present work, three different q-RASPR models (MLR q-RASPR using the same number of descriptors as that of the QSPR, PLS q-RASPR with the same number of descriptors as that of the QSPR, and an univariate q-RASPR) were developed against each dataset for a comparative statistical analysis between QSPR and q-RASPR models

and to find the best one for prediction. The complete model statistics of the PLS q-RASPR models of all eight datasets have been shown in **Table 3,** and the complete analysis has been presented in **Supplementary Material 1**. Additionally, we have reported the univariate q-RASPR model statistics in **Table 3**, while the univariate models of the individual local datasets can be found in **Supplementary Material 1**. The univariate q-RASPR models of Global datasets have performed better in test set predictions which is evident from the external validation metrics given in **Table 3**.

Table 3. Validation statistics of the developed PLS q-RASPR models and global univariate models

Datasets	Model	nDesc	LVs	F-value	R^2	Q^2_{LOO}	Q^2_{F1}	Q^2_{F2}	MAE_{Test}
Dataset 1	PLS q-RASPR	4	1	437.538	0.948	0.934	0.949	0.949	0.299
Dataset 2	PLS q-RASPR	6	3	167.111	0.94	0.922	0.949	0.948	0.239
Dataset 3	PLS q-RASPR	2	1	274.909	0.945	0.934	0.941	0.941	0.428
Dataset 4	PLS q-RASPR	4	3	223.218	0.971	0.958	0.893	0.893	0.44
Dataset 5	PLS q-RASPR	3	1	183.000	0.915	0.877	0.899	0.89	0.638
Global Dataset 1	PLS q-RASPR	7	3	752.143	0.972	0.961	0.927	0.926	0.392
	Univariate q-RASPR	1	-	712.070	0.914	0.908	0.957	0.956	0.316
Global Dataset 2	PLS q-RASPR	7	6	223.043	0.953	0.943	0.901	0.839	0.369
	Univariate q-RASPR	1	-	475.154	0.870	0.861	0.930	0.887	0.309
Global Dataset 3	PLS q-RASPR	6	2	624.408	0.913	0.906	0.938	0.938	0.415
	Univariate q-RASPR	1	-	855.610	0.877	0.872	0.933	0.933	0.398

nDesc = the number of descriptors, LVs = the number of latent variables

The RASPR descriptors are a set of similarity and error-based measures for a particular query compound. Unlike standard QSPR descriptors which describe the structural and physicochemical nature of a query compound, the RASPR descriptors do not describe the nature of that particular query compound and instead demonstrate information of its close congeners from the source set. The list of RASPR descriptors appearing in our models is presented in **Table 4**.

Table 4. List of RASPR descriptors and their significance

RASPR descriptors	***Significance***	***Occurrences in the models***
RA function	A Read-Across-derived composite function encoding information of all the selected structural and physicochemical descriptors	All the datasets
SD Similarity	Standard deviation of the similarity values of the close congeners for a particular query compound	Dataset 1
MaxPos	Similarity value to the closest positive close source compound for a particular query compound	Dataset 2
g_m**Avg.Sim*	A derived RASPR descriptor obtained by the product of g_m and *Avg.Sim*	Dataset 2, Global Dataset 1
s_m^2	Banerjee-Roy similarity coefficient 2	Dataset 3
Neg.Avg.Sim	Average similarity value of the negative close source compounds for each query compound	Dataset 4
s_m^1	Banerjee-Roy similarity coefficient 1	Dataset 4
g_m*_class*	A derived descriptor from gm having values 0 and 1	Dataset 4
SD Activity	Weighted standard deviation of the response values of the close source compounds for each query compound	Dataset 5
CVact	Coefficient of variation of the response values of the close source compounds for each query compounds	Global Dataset 2
CVsim	Coefficient of variation of the similarity values of the close source compounds for each query compound	Global Dataset 2

3.3 Insights of the molecular descriptors identified by the PLS QSPR models

The mechanism of adsorption of molecules onto microplastics is indeed a critical aspect to consider. The interaction between microplastics and chemical compounds occurs through various mechanisms, including physical adsorption, surface complexation, and partitioning. These mechanisms are influenced by the properties of both the microplastic particles and the compounds themselves. Techniques such as batch adsorption experiments, surface analysis (e.g., FTIR, XPS), and modeling approaches help elucidate the factors influencing adsorption mechanism. Among various factors, surface area, surface chemistry, porosity, hydrophobicity, size, and shape of microplastic particles affect the adsorption of contaminants onto MPs (52). Microplastics have a large surface area-to-volume ratio, providing ample sites for adsorption of chemical compounds. Porous MPs can provide additional sites for adsorption due to increased surface area. The surface chemistry of MPs including functional groups and charges, can influence the adsorption of different types of molecules, e.g., hydrophobic MPs tend to adsorb hydrophobic compounds more readily due to favourable hydrophobic interactions. The size and shape of MPs can also affect the accessibility of adsorption sites and the extent of interaction with chemical compounds. Likewise, the hydrophobicity, molecular size, polarity, and chemical structure of the organic contaminants can affect the adsorption onto the surface of microplastics. The adsorptive attributes of plastic particles and their associated chemistry are influenced by weathering processes also. These processes impact the transport, fate, and toxicity of MPs and other environmental contaminants. In the case of adsorption, chemical contaminants adhere to the solid surfaces of MPs and NPs through physical attachment. The adsorption mechanism of chemical contaminants on plastic particles depends on various physicochemical and environmental factors, including the type and characteristics of pollutants, free energy relationships, and the textural and surface chemistry of the microplastic/nanoplastics in solid phase. In this section, we discuss the molecular descriptors

identified by the three Global PLS QSPR models. Although the univariate q-RASPR models are better predictive than the corresponding QSPR, the single variable *RAfunction* has been extracted from the information of molecular descriptors of corresponding PLS QSPR models only. Thus, these identified features portray the impact of structures of organic contaminants for their adsorption onto microplastics in the unspecified aqueous environment.

3.3.1 Global Model 1

The PLS QSPR model of Global dataset 1 is developed with five molecular descriptors (BLTD48, C%, F08[Cl-Cl], B10[C-F], and F07[C-C]) and two indicator variables (Ind1(pe), and Ind2(pp)). We have constructed the Loading plot (see **Figure 4**) of the PLS QSPR model using Simca-P software to get an idea of the relative importance of modeled descriptors and their contribution. The loading plot is usually constructed with the X- and Y-variables that get projected on the 1st two latent variable spaces. The X-variables are placed along the 1st LV based on their variable influence on projection (VIP) values and thus, the variables that are placed far from the origin play most significant role. Here, C% and BLTD48 are the farthest from the origin and hence most significant descriptors of Global dataset 1. C% is placed near the Y-variable (log Kd) and in the same quadrant that signifies the positive contribution of it. On the other hand, BLTD48 is placed in the opposite quadrant of Y-variable which signifies its negative contribution. Among the rest, F08[Cl-Cl], Ind1(pe), and B10[C-F] are moderately influential while, F07[C-C] and Ind2(pp) are the least influential descriptors. We have tabulated the descriptors of Global Model 1 and their significance in **Table S2 of Supplementary Material 2**.

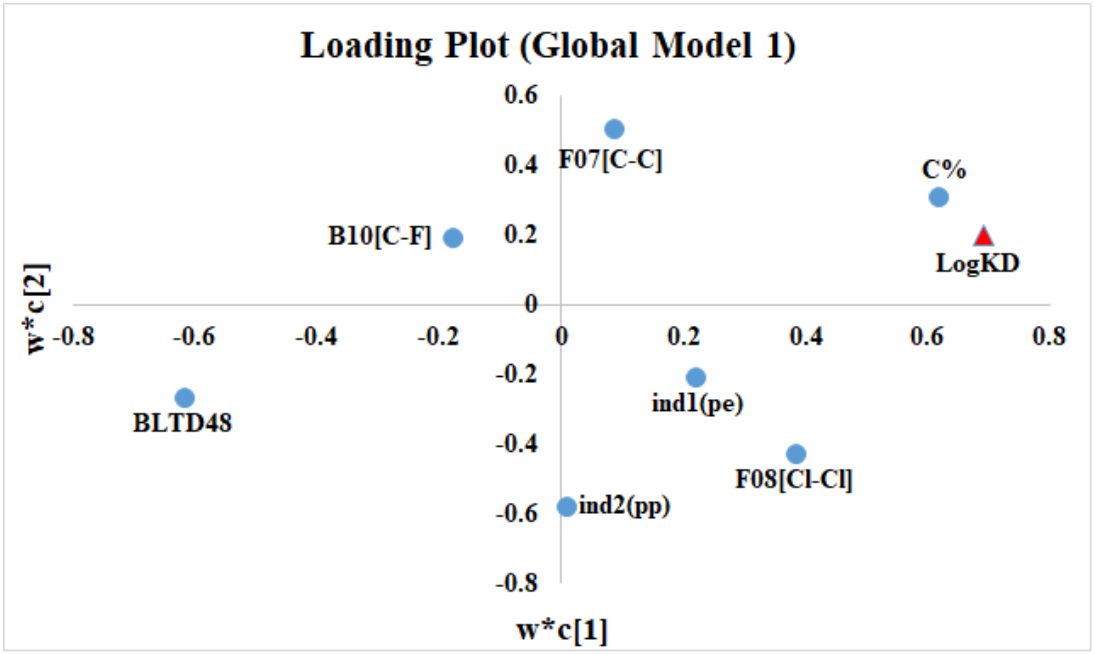


Figure 4 Loading plot of Global model 1 (PLS QSPR)

3.3.2 Global Model 2

The PLS QSPR model of Global dataset 2 is developed with seven molecular descriptors namely MLOGP2, F08[Cl-Cl], SaaaC, F07[O-F]), H-047, nHet, and B03[N-O]. From the Loading plot presented in **Figure 5**, MLOGP2 and F08[Cl-Cl] are found to be the most significant descriptors of Global Model 2 (PLS QSPR). Both of these descriptors have been

found in proximity to the Y-variable (log Kd) which indicates their positive contribution. B03[N-O] is a descriptor with moderate influence and the rest are least important descriptors of the model. We have tabulated the descriptors of Global Model 2 and their significance in **Table S2 of Supplementary Material 2**.

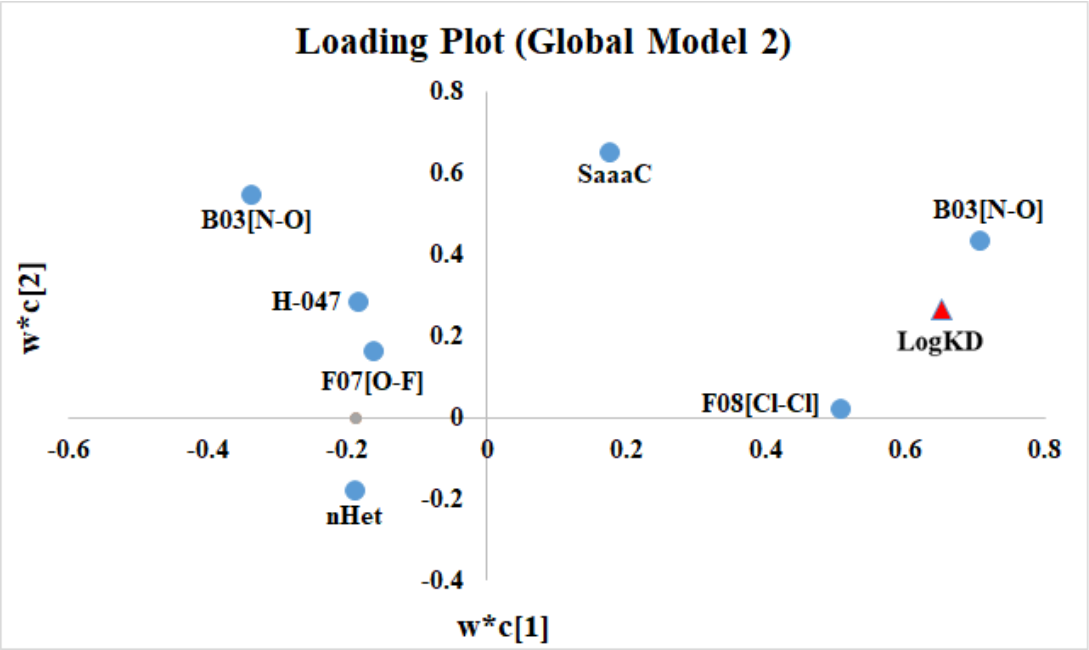


Figure 5 Loading plot for Global model 2 (PLS QSPR)

3.3.3 Global Model 3

The PLS QSPR model of Global dataset 3 is developed with six molecular descriptors namely MLOGP2, F08[Cl-Cl], F07[C-N], SaaaC, nHet, and one indicator variable Ind2(pp). From the Loading plot presented in **Figure 6**, MLOGP2 and F07[C-N] are found to be the most significant descriptors of Global Model 3 (PLS QSPR). MLOGP2 is placed near the Y-variable (log Kd) and in the same quadrant which signifies the positive contribution of it. On the other hand, F07[C-N] is placed in the opposite quadrant of Y-variable which signifies its negative contribution. F08[Cl-Cl] and nHet are the descriptors with moderate influence while SaaaC and Ind2(pp) are the least important descriptors of the model. The number of heteroatoms, hydrophobic bulk, hydrophobic interaction, percentage of C atom, carbon skeleton of compounds, and van der Waals interactions collectively determine the adsorption capacity of microplastics in a given environment. We have tabulated the descriptors of Global Model 3 and their significance in **Table S2 of Supplementary Material 2**.

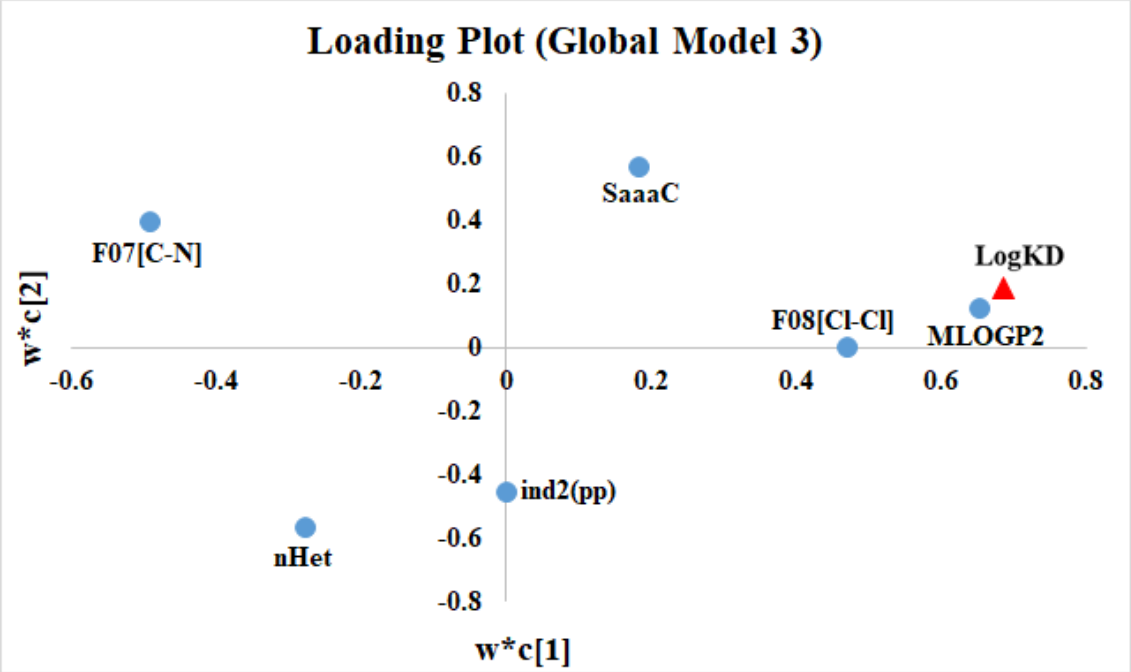


Figure 6 Loading plot for Global model 3 (PLS QSPR)

4. Statistical analysis of the results from PLS QSPR and PLS q-RASPR models

A common obstacle faced by the QSPR modelers is the lack of experimental property/toxicity data for a particular endpoint, which hinders the reliability of the developed models. The rule of thumb for developing a statistically reliable model is that it should be trained using a higher number of compounds with a lower number of descriptors that should encode the complete chemical space. This increases the degree of freedom of the statistical model, resulting in reliable predictions for external query compounds. However, this is not always the case since many times the modeler may require a higher number of descriptors to encode the complete and complex chemical space. Researchers have tried to address this problem by adopting non-statistical aspects like read-across. Although this approach generates reliable predictions, the quantification of the contributions of the features is a major concern. Recently, Banerjee et al. (45) demonstrated a statistical modeling algorithm (q-RASPR) applied to a small dataset relating to the cytotoxicity of hybrid TiO_2 nanoparticles and showed that this modeling approach can generate reliable predictions even in case of limited data points. Using this idea, we have statistically evaluated the reliability of the q-RASPR modeling approach and showed that although this is statistically driven, it not only generates reliable predictions with enhanced predictivity but also increases the significance of the regression coefficients by using a lower number of descriptors and latent variables.

For the efficient and effective comparison, we have compared the QSPR and q-RASPR model statistics by not only using a common modeling algorithm – the Partial Least Squares (PLS), but also using the same number of modeling descriptors. Additionally, we have generated univariate q-RASPR models that show good model statistics for further exploration. In order to ensure that chemical information is not lost in the development of PLS q-RASPR models (since previous studies showed that Multiple Linear Regression (MLR) models are best suited for q-RASPR modeling (45,53) demonstrating the absence of significant inter-correlations), additional MLR q-RASPR models have also been developed using the same descriptor combinations used in the PLS q-RASPR models. The above-mentioned types of models have

been developed for all five individual datasets and the three global datasets. The chemical space encoded by all the models has been evaluated by identifying the structural outliers using the leverage approach. (54) The *F*-values of all the models were also computed to understand the significance of the determination coefficients of the developed models.

Analyzing the model statistics of PLS QSPR, PLS q-RASPR, and univariate q-RASPR for all the datasets (**Tables 2** and **3**), it was observed that in 6 out of the 8 datasets, the q-RASPR models have higher *F*-values. This can be justified since the univariate q-RASPR models generate very good results, and also the PLS q-RASPR models used a lower number of latent variables. In terms of the predictive ability, it was observed that in 6 out of the 8 datasets, the MAE_{test} value is lower in the case of q-RASPR models. The most interesting observation in this aspect is that in all 6 cases, the univariate q-RASPR model had the lowest MAE_{test} values. Additionally, among these 6 cases, the PLS q-RASPR models had lower MAE_{test} values than the corresponding PLS QSPR models. It is essential to note that the MLR q-RASPR models have not been taken into consideration for the comparison, since we are only focused on comparing QSPR and q-RASPR using the same modeling algorithm and the same number of descriptors and also evaluated the potential performance of a univariate q-RASPR model. As evident from the previous studies (55, 56), the q-RASPR models have a higher predictivity when they are developed using a lower number of descriptors. However, for efficient comparison purposes, we have developed q-RASPR models using the same number of descriptors as that of the corresponding QSPR model. Analyzing the chemical space that the models encode, the number of structural outliers of the training and test sets were determined using the leverage approach. From this observation, it was found that in most of the cases, the q-RASPR modeling approach generated lower structural outliers. Here also, the univariate q-RASPR model played a great role in encapsulating the entire chemical space into a single descriptor, thus generating lower outlier counts for the training and test sets. **Figure 7** shows the comparative performance of q-RASPR and QSPR global models The complete comparison table, representing the number of descriptors, latent variables, internal and external model statistics, the number of structural outliers for the training and test sets and model-specific inference has been provided in an Excel sheet of **Supplementary Material 1.** From the collective observations, it can be concluded that q-RASPR models generate enhanced predictivity, increased reliability in predictions and are statistically more significant as compared to conventional QSPR models.

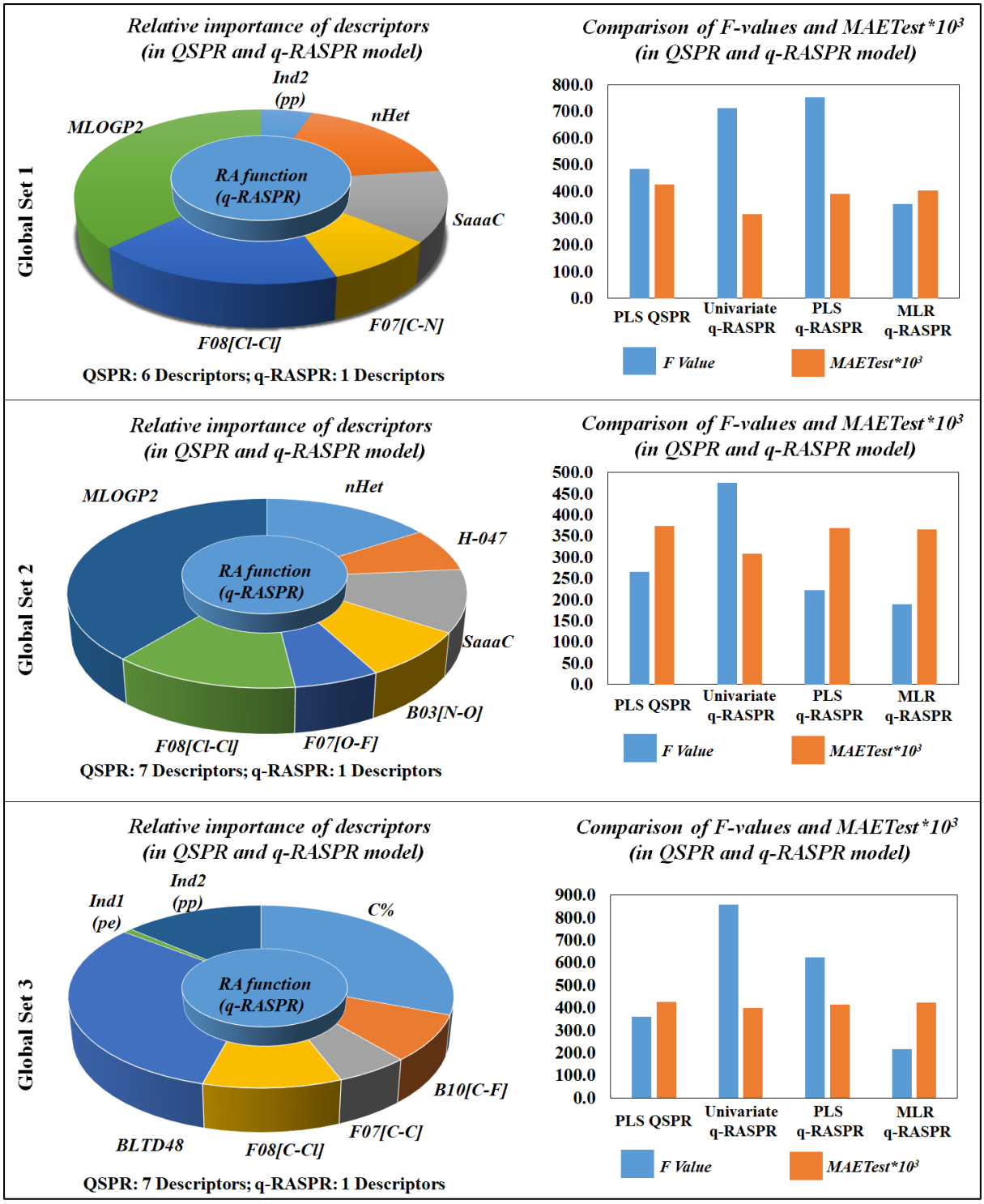


Figure 7. Comparison of the quality of QSPR and q-RASPR global models on the basis of number of descriptors, F value and MAE_{Test}.

5. Comparison with previous work

In the previous work, Li et al. (2020) (28) used a set of 3D descriptors to develop their QSPR model while we have computed only 2D descriptors used for model development which do not require prior structural optimization. On comparing the errors for the test sets with the previous models, it was found that our PLS q-RASPR models show lower errors in terms of MAE_{test} in most of the cases. Earlier the authors (28) had developed five different QSPR models for each type of microplastic in different water environments, but in this work, we have developed five

local q-RASPR models and three univariate q-RASPR global models. All global models developed considering different types of microplastics and different aqueous environments show high $Q^2_{F1\ and}$ Q^2_{F2} values and low MAE values. **Table 5** compares the MAE values between our models and the models reported earlier.

Table 5. A comparison of MAE values between our model and the models reported earlier (28)

Dataset	**MAE_{test}**	
	q-RASPR (PLS) (Present work)	**QSPR (Previous work) (28)**
Data set 1	0.299	0.664
Data set 2	0.239	0.384
Data set 3	0.428	0.467
Data set 4	0.441	0.228
Data set 5	0.638	0.654
	q-RASPR (Univariate) (Present work)	**QSPR (Previous work)**
Global set 1	0.316	-
Global set 2	0.309	-
Global set 3 (whole data set)	0.398	-

6. Limitations

For the modeling of microplastics' adsorption capacity for organic contaminants in diverse aqueous environments, 162 data points were used. The local q-RASPR models were developed with a comparatively smaller number of data points than the global models. To achieve greater statistical reliability, large datasets are always preferred. However, due to the scarcity of experimental data, a comparatively small number of data points have been used in the local model development. Besides this, only three types of microplastics (polyethene, polypropylene, and polystyrene) were taken into account in this study. There are several other classes of microplastics, i.e., polyvinyl chloride, poly ethyl terephthalate, polybrominated diphenyl ethers, etc. that can also be considered. So there is still a significant scope for improvement of the existing models by expanding the chemical diversity of the training data in future work.

7. Conclusion

The adsorption of organic pollutants onto microplastics is a critical environmental issue with far-reaching implications for the ecosystem's health and human well-being. The q-RASPR approach represents a significant advancement in our ability to predict and understand the adsorption behavior of these pollutants on microplastics. By establishing quantitative relationships between the chemical structures of pollutants and their adsorption behavior, these models provide valuable insights into the key molecular properties such as the number of heteroatoms, hydrophobic bulk, hydrophobic interaction, percentage of C atom, carbon skeleton of compounds, and van der Waals interactions collectively determine the adsorption capacity of microplastics in a given environment and also the chemical composition, size, and

shape of the MPs/ NPs that influence the adsorption process. The statistical results and applicability domain analysis of the q-RASPR models indicate the satisfactory goodness-of-fit, robustness, and predictive ability of these models superseding the quality of QSPR models and most of the previously reported models on this data set. Besides these, the present research mainly focused on developing a global q-RASPR model that can efficiently predict microplastics' adsorption capacity (Kd) for various organic pollutants after considering the aqueous environment's versatility or the microplastic's chemical nature. The q-RASPR models developed from the Global datasets in the present study are clearly more advantageous in terms of versatility than the previously developed local models. The efficient predictions obtained from the global models can help in regulatory decision-making, guide pollution control measures, and contribute to the sustainable management of plastic waste and environmental contamination. Furthermore, the q-RASPR models offer a rapid and efficient way to screen a large number of chemicals, enabling researchers to prioritize compounds for further experimental studies. By leveraging data from structurally similar compounds, this approach enhances our ability to predict the adsorption behavior of new chemicals with limited experimental data. Overall, the development of QSPR and q-RASPR models for the adsorption of organic pollutants onto microplastics represents a promising tool for addressing this complex environmental challenge. These models have the potential to drive advancements in pollution control strategies, support evidence-based decision-making, and ultimately contribute to a healthier and more sustainable environment for future generations.

Supplementary Information Available
Supplementary Material 1 contains the raw data used for the development of the models.
Supplementary Material 2 contains Supplementary Tables and Figures.

Acknowledgment
AB thanks LSRB, DRDO, New Delhi for a Senior Research Fellowship.

Funding
This research is funded by the University Grants Commission (UGC), New Delhi, India

Declaration of conflict of interest
None

References

(1) M. Enfrin, L. F. Dumee, and J. Lee, Nano/microplastics in water and wastewater treatment processes – origin, impact and potential solutions, *Water Res.*, 2019, **161**, 621–638.

(2) S. Borrelle, J. Ringma, K. Law, C. Monnahan, L. Lebreton, A. McGivern, E. Murphy, J. Jambeck, G. Leonard, M. Hilleary, M. Eriksen, H. Possingham, H. De Frond, L. Gerber, B. Polidoro, A. Tahir, M. Bernard, N. Mallos, M. Barnes, and C. Rochman, Predicted growth in plastic waste exceeds efforts to mitigate plastic pollution, *Science,* 2020, **369**, 1515–1518.

(3) R. Geyer, J. R. Jambeck, and K.L.Law, Production, use, and fate of all plastics ever made, *Sci. Adv.,* 2017, **3**, 1700782.

(4) J. Li, Y. Song, and Y. Cai, Focus topics on microplastics in soil: Analytical methods, occurrence, transport, and ecological risks, *Environ. Pollut.,* 2020, **257**, 113570.

(5) L. Nizzetto, G. Bussi, M. N. Futter, D. Butterfield, and P. G. Whitehead, A theoretical assessment of microplastic transport in river catchments and their retention by soils and river sediments, *Environ.Sci.Proc.,* 2016, **18**,10501059.

(6) C. J. Tien, Z.-X. Wang, and C. S. Chen, Microplastics in water, sediment and fish from the Fengshan River system: relationship to aquatic factors and accumulation of polycyclic aromatic hydrocarbons by fish, *Environ. Pollut.,* 2020, **265**, 114962.

(7) S. Sharma, and S. Chatterjee, Microplastic pollution, a threat to marine ecosystem and human health: a short review, *Environ. Sci. Pollut. Res.,* 2017, **24**, 21530–21547.

(8) M. Cole, P. Lindeque, C. Halsband, and T.S. Galloway, Microplastics as contaminants in the marine environment: are view, *Mar. Pollut.,* 2011, **62**, 2588–2597.

(9) C.G. Avio, S. Gorbi, M. Milan, M. Benedetti, D. Fattorini, G. d'Errico, M. Pauletto, L. Bargelloni, and F. Regoli, Pollutants bioavailability and toxicological risk from microplastics to marine mussels*, Environ. Pollut.,* 2015,**198**, 211–222.

(10) B. Zhao, P. Rehati, Z. Yang, Z. Cai, C. Guo, and Y. Li, The potential toxicity of microplastics on human health, *Sci. Total Environ.*, 2023, **912**, 168946.

(11) Y. Li, L. Tao, Q. Wang, F. Wang, G. Li, and M. Song, Potential health impact of microplastics: a review of environmental distribution, human exposure, and toxic effects. Environ. Health. 2023, **1**, 249-57.

(12) T. Braun, L. Ehrlich, W. Henrich, S. Koeppel, I. Lomako, P. Schwabl, and B. Liebmann. Detection of microplastic in human placenta and meconium in a clinical setting, *Pharmaceutics*, 2021, **13,** 921.

(13) L. Zhu, J. Zhu, R. Zuo, Q. Xu, Y. Qian, and A.N. Lihui, Identification of microplastics in human placenta using laser direct infrared spectroscopy, *Sci. Total Environ*. 2023, **856**, 159060.

(14) R. Marfella, F. Prattichizzo, C. Sardu, G. Fulgenzi, L. Graciotti, T. Spadoni, N. D'Onofrio, L. Scisciola, R. La Grotta, C. Frigé, and V. Pellegrini. Microplastics and Nanoplastics in Atheromas and Cardiovascular Events, *N. Engl. J. Med*., 2024, **390**, 900-910.

(15) H. Lee, W. J. Shim, and J.-H. Kwon, Sorption capacity of plastic debris for hydrophobic organic chemicals, *Sci. Total Environ.,* 2014, **470–471**, 1545–1552.

(16) A. Elizalde-Velázquez, S. Subbiah, T. A. Anderson, M. J. Green, X. Zhao, and J. E. Cañas-Carrell, Sorption of three common nonsteroidal anti-inflammatory drugs (NSAIDs) to microplastics*, Sci. Total Environ.,* 2020*,***715**, 136974.

(17) J. Li, K. N. Zhang, and H. Zhang, Adsorption of antibiotics on microplastics, *Environ. Pollut. ,*2018, **237**, 460–467.

(18) Y. Mato, et al. Plastic resin pellets as a transport medium for toxic chemicals in the marine environment, *Environ. Sci. Technol.,*2001, **35**,318–324.

(19) A.S Delle. Factors affecting sorption of organic compounds in natural sorbent/water systems and sorption coefficients for selected pollutants. A review. *J. Phys. Chem. Ref. Data* . 2001, **30**, 187-439.

(20) J.C. Madden, S.J. Enoch, A. Paini, and M.T. Cronin. A review of in silico tools as alternatives to animal testing: principles, resources and applications. *Alter. Lab. Animals*. 2020 **48**, 146-72.

(21) A. R. Katritzky, V. S. Lobanov and M. Karelson, QSPR: the correlation and quantitative prediction of chemical and physical properties from structure, *Chem. Soc. Rev*., 1995, **24**, 279–287.

(22) N. Ball, M.T. Cronin, J. Shen, K. Blackburn, E.D. Booth, M. Bouhifd, E. Donley, L. Egnash, C. Hastings, D.R. Juberg, and A. Kleensang . T4 report: Toward good read-across practice (GRAP) guidance. *Altex.*, 2016, **33**, 149.

(23) G. Patlewicz, M.T. Cronin, G. Helman, J.C. Lambert, L.E. Lizarraga, and I. Shah . Navigating through the minefield of read-across frameworks: A commentary perspective. *Comput. Toxicol.*, 2018, **6**, 39-54.

(24) S. Manganelli and E. Benfenati, in Use of read-across tools, *In Silico* Methods for Predicting Drug Toxicity, ed. E. Benfenati, Humana Press, 2016, pp. 305–322.

(25) A. Banerjee, A. Gajewicz-Skretna, and K. Roy. A machine learning q-RASPR approach for efficient predictions of the specific surface area of perovskites, *Mol. Inform*., 2023 **42**, 2200261.

(26) A. Banerjee and K. Roy. First report of q-RASAR modeling toward an approach of easy interpretability and efficient transferability, *Mol. Divers*., 2022, **26**, 2847-2862.

(27) Y. Zhao, J.W. Choi, J.K. Bediako, M.H. Song, S. Lin, C.W Cho, and Y.S. Yun, Adsorptive interaction of cationic pharmaceuticals on activated charcoal: experimental determination and QSAR modelling, *J. Hazard. Mater*., 2018, **360**, 529-535.

(28) M. Li, H. Yu, Y. Wang, J. Li, G. Ma, and X. Wei. QSPR models for predicting the adsorption capacity for microplastics of polyethylene, polypropylene and polystyrene, *Sci. Rep*., 2020, **10**, 14597.

(29) R. Todeschini and V. Consonni, Handbook of Molecular Descriptors, John Wiley & Sons, Germany, 2008

(30) A. Mauri, in alvaDesc: a tool to calculate and analyse molecular descriptors and fingerprints, Ecotoxicological QSARs. Methods in Pharmacology and Toxicology, ed. K. Roy, Humana, 2020, pp. 801–820.

(31) M. Chatterjee, A. Banerjee, S. Tosi, E. Carnesecchi, E. Benfenati, and K. Roy. Machine learning-based q-RASAR modeling to predict acute contact toxicity of binary organic pesticide mixtures in honey bees, *J. Hazard. Mater*., 2023, **460**, 132358.

(32) J.G. Krishna, P.K. Ojha, S. Kar, K. Roy, and J. Leszczynski, Chemometric modeling of power conversion efficiency of organic dyes in dye sensitized solar cells for the future renewable energy, *Nano Energy*, 2020, **70**, 104537

(33) S. Ghosh, M. Chatterjee, and K. Roy, Predictive quantitative read-across structure–property relationship modeling of the retention time (Log t R) of pesticide residues presents in foods and vegetables, *J. Agric. Food Chem*., 2023, **71**, 9538-9548.

(34) J. T. Leonard, and K. Roy. "On selection of training and test sets for the development of predictive QSAR models, *QSAR Comb. Sci.*, 2006. **25**, 235-251.

(35), T. Yu, G. Yu, P. Y. Li, and L. Wang, Citation impact prediction for scientific papers using stepwise regression analysis, *Scientometrics*, 2014, **101**, 1233-1252.

(36) P. K. Ojha, and K. Roy, Comparative QSARs for antimalarial endochins: importance of descriptor-thinning and noise reduction prior to feature selection. *Chemom. Intll. Lab. Syst.,* 2011, **109**, 146-161.

(37,) J. Roy, S. Ghosh, P. K. Ojha, and K. Roy, Predictive quantitative structure–property relationship (QSPR) modeling for adsorption of organic pollutants by carbon nanotubes (CNTs), *Environ. Sci. Nano*, 2019, **6**, 224-247.

(38) L. Jiao, and H. Li, QSPR studies on the aqueous solubility of PCDD/Fs by using artificial neural network combined with stepwise regression. *Chemom. Intll. Lab. Syst*., 2010, **103**, 90-95.

(39) S. Pirhadi, F. Shiri, and J. B. Ghasemi, Multivariate statistical analysis methods in QSAR, *Rsc Advances*, 2015, **5**, 104635-104665.

(40) K. Roy, On some aspects of validation of predictive quantitative structure–activity relationship models, *Expet Opin. Drug Discovery*, 2007, **2**, 1567–1577

(41) S. Wold, M. Sjöström, and L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intll. Lab. Syst., 2001, **58**, 109-130.

(42) M. Chatterjee, A. Banerjee, P. De, A. Gajewicz-Skretna and K. Roy, a novel quantitative read-across tool designed purposefully to fill the existing gaps in Nano safety data, *Environ. Sci. Nano*, 2022, **9**, 189–203.

(43) A. Banerjee, M. Chatterjee, P. De and K. Roy, Quantitative predictions from chemical read-across and their confidence measures, *Chemom. Intell. Lab. Syst.,* 2022, **227**,104613.

(44) S. Manganelli and E. Benfenati, Use of read-across tools, In Silico Methods for Predicting Drug Toxicity, ed. Benfenati E., Springer, Milan, Italy, 2016, pp. 305–322.

(45) A. Banerjee, S. Kar, S. Pore, and K. Roy. Efficient predictions of cytotoxicity of TiO2-based multi-component nanoparticles using a machine learning-based q-RASAR approach, *Nanotoxicology*, 2023, **17**, 78-93.

(46) M. Goodarzi, B. Dejaegher and Y. V. Heyden, Feature selection methods in QSAR studies, *J. AOAC Int.,* 2012, **95**, 636–651.

(47) M. Shahlaei, Descriptor selection methods in quantitative structure–activity relationship studies: a review study, *Chem. Rev.,* 2013, **113**, 8093–8103.

(48) K. Roy, S. Kar and R. N. Das, Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment, Academic Press, NY, 2015.

(49) K. Roy, R. N. Das, P. Ambure and R. B. Aher, Be aware of error measures. Further studies on validation of predictive QSAR models, *Chemom. Intell. Lab. Syst*., 2016, **152**, 18–33.

(50) K. Roy, S. Kar and R. N. Das, A Primer on QSAR/QSPR Modeling: Fundamental Concepts, Springer, 2015, pp. 45–46.

(51) K. Roy, S. Kar and P. Ambure, on a simple approach for determining applicability domain of QSAR models, *Chemom. Intell. Lab. Syst*., 2015, **145**, 22–29.

(52) P. K. Rai, C. Sonne, R.J.C. Brown, S. A. Younis, and K.-H. Kim, Adsorption of environmental contaminants on micro- and nano-scale plastic polymers and the influence of weathering processes on their adsorptive attributes, *J. Hazard.Mater*. 2022, **427**, 127903.

(53) A. Banerjee and K. Roy. On some novel similarity-based functions used in the ML-based q-RASAR approach for efficient quantitative predictions of selected toxicity end points. *Chem. Res. Toxicol*. 2023, **36**, 446-464.

(54) P. Gramatica. Principles of QSAR models validation: internal and external, *QSAR Comb. Sci*., 2007, **26**, 694-701.

(55) K. Roy and A. Banerjee, q-RASAR: A Path to Predictive Cheminformatics. Springer, NY, 2024.

(56) A. Banerjee and K. Roy, How to correctly develop q-RASAR models for predictive cheminformatics. *Expert Opin. Drug Discov.* 2024, https://doi.org/10.1080/17460441.2024.2376651 .