Predicting the dissolution of metal-based nanoparticles by means of QSPRs and the effect of data augmentation

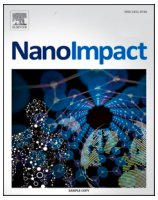
Song, Y.; Balraadjsing, S.; Peijnenburg, W.J.G.M.; Vijver, M.G.

NanoImpact 37 (2025) 100547

Contents lists available at ScienceDirect

NanoImpact

journal homepage: www.elsevier.com/locate/nanoimpact



Predicting the dissolution of metal-based nanoparticles by means of QSPRs and the effect of data augmentation

Yuchao Song [a,1], Surendra Balraadjsing [a,*,1], Willie J.G.M. Peijnenburg [a,b], Martina G. Vijver [a]

[a] Institute of Environmental Sciences (CML), Leiden University, Einsteinweg 2, 2333 CC Leiden, the Netherlands
[b] National Institute of Public Health and the Environment (RIVM), Center for Safety of Substances and Products, P.O. Box 1, Bilthoven, the Netherlands

ARTICLE INFO

Editor: Bernd Nowack

Keywords:
Metallic nanoparticles
Fate dynamics
Quantitative structure-property model
Machine learning
Dissolution

ABSTRACT

Particle dissolution is a critical process in the environmental fate assessment of metal-based nanoparticles (MNPs). Numerous attempts have been made previously to adequately quantify dissolution (kinetics), however, existing dissolution data and models are generally limited to a few nanomaterials or specific time points. Hence, they only capture phases of the process. This study aimed to develop a Quantitative Structure-Property Relationship (QSPR) model to predict the ion release (in %) of MNPs for different time points and water chemistry conditions. Furthermore, many machine learning models are frequently plagued by a lack of data and recently data augmentation has been suggested as a method to mitigate this issue. Therefore, we also investigated the effects of data augmentation on QSPRs. Following data collection from literature, QSPR models were generated and results indicate models with adequate performance ($R^2 > 0.7$). Results also demonstrated significant improvements in model performance with increasing amounts of applied data augmentation. However, a deeper evaluation of the results also highlighted that data augmentation can lead to misleading and overoptimistic model evaluation. Thus, proper model assessment is necessary when evaluating QSPRs. Variable importance analysis results revealed that the "initial concentration" and features related to the size and shape of MNPs were the most critical factors in the dissolution process. The predictive models generated here for MNP dissolution can improve nanomaterial testing efficiency and guide experimental design.

1. Introduction

Metal-based nanoparticles (MNPs) have been increasingly implemented into household (Lowry et al., 2019) and consumer products (J. Li et al., 2022) over the past decades. Following the use of such a product, a wide range of structurally and chemically diverse MNPs will unavoidably be released into the aquatic environment (Gardea-Torresdey et al., 2014; Poynton et al., 2019). When released into aquatic ecosystems, MNPs undergo various transformations like dissolution, aggregation/agglomeration, adsorption, deposition and redox reactions (Amde et al., 2017; Hedberg et al., 2019; Tangaa et al., 2016). Particle dissolution is considered a key parameter that mediates MNP fate and toxicity. After MNPs dissolve, they will predominantly be present in two forms: MNP particles and metal ions shed from the MNPs (X. Li et al., 2010). These different forms influence the behavior and distribution of the MNPs across environmental compartments. This in turn also influences the exposure potency as well as potential toxic effects of MNPs towards aquatic species (Ding et al., 2022; Kang and Park, 2021; Tschiche et al., 2022). In general, – depending on the dissolution rate - both particulate and ionic forms of MNPs contribute significantly to overall toxicity (Tripathy et al., 2014; Zhai et al., 2017). The ion release is influenced by several factors, including the physico-chemical properties of MNPs (*e.g.* particle size, shape, surface chemistry) and the composition of the surrounding environment (*e.g.* pH, temperature, ionic strength) (Loza et al., 2014).

The wide variety of MNPs with different physico-chemical properties makes it impossible to measure all processes at the particle surface, especially when they are exposed under different conditions (Odzak et al., 2015) and hence fate dynamics can fluctuate significantly. To address this problem, developing computational models to predict the properties or toxic effects of MNPs in a prospective manner without testing all combinations and dynamics, is urgently needed in view of the complexity of MNPs and their interactions within aquatic environments. Among the new approach methodologies (NAMs) based on *in silico*

* Corresponding author.
E-mail address: s.balraadjsing@cml.leidenuniv.nl (S. Balraadjsing).
[1] Contributed equally to this work

https://doi.org/10.1016/j.impact.2025.100547
Received 22 November 2024; Received in revised form 15 January 2025; Accepted 30 January 2025
Available online 7 February 2025

predictions, Quantitative Structure-Property Relationship (QSPR) modelling proves to be a useful tool for the prediction of material properties by providing a mathematical correlation with its chemical structural features (J. Li et al., 2022). To date, various QSPR models have been created in relation to MNPs with the majority focusing on endpoints such as zeta potential (Abbas et al., 2020; Mikolajczyk et al., 2015; Toropov et al., 2018). While particle dissolution is considered a crucial element of MNPs, QSPRs for the prediction of particle dissolution remain explored to a lesser extent. Recently, few QSPR methods have been applied in methodological studies for the quantitative prediction of particle solubility, however these are mainly limited to C_{60} and C_{70} fullerenes (Gupta and Basant, 2018; Toropova et al., 2011). Another aspect that is often also overlooked in QSPRs (and QSARs) is the influence of exposure conditions (Balraadjsing et al., 2024). By incorporating a wide range of MNP descriptors (*e.g.* size, shape, surface chemistry, and surface area-to-volume ratio), along with experimental dissolution data, QSPR models have the potential to effectively predict particle dissolution and provide valuable insights into the factors influencing MNP dissolution. This also enables safer and more efficient design of MNPs and their application in various fields.

Meanwhile, machine learning (ML) enables researchers to analyze large data sets by training models that can be used to classify observations into discrete groups, learning which features are important, or predicting the outcome of new experiments (Brown et al., 2020). ML algorithms such as random forests, support vector machines, and neural networks, have shown great promise in handling complex data sets and extracting meaningful patterns from toxicity data (Huang et al., 2022).

The main goal of this study is to develop a ML-based QSPR model that can predict the percentage of ion release from MNPs at any time point during the dissolution process. To achieve this goal, we performed two research steps. The first step was collecting data on MNPs physico-chemical and water chemistry properties using published literature where MNP dissolution was measured experimentally. As a second step, data amplification was used to gradually increase the size of the collected data after which QSPR models were created and variable importance was assessed. By gradually increasing the amount of amplified data, we could also investigate the effect data amplification has on model performance. The QSPR model developed here helps gain insight into the dynamic ion release process of MNPs and the necessary factors needed to create such QSPR models. By also investigating the effect of data augmentation, we gain deeper knowledge into how machine learning models are affected by it and whether this can be a feasible option for when toxicity data is scarce.

2. Methods

2.1. General overview

An ecotoxicity dataset previously used by Balraadjsing et al. (Balraadjsing et al., 2024) was used and modified with the focus on particle dissolution rather than toxicity. After modifying the data set, data augmentation was applied to artificially increase the data set, resulting in five different data sets with varying sizes. Subsequently, QSPR models were created based on the five data sets and their performance were evaluated. Additionally, the model performance was also evaluated on "novel" materials, not included during the training of the models. Finally, the uncertainty of the models (through the applicability domain and prediction intervals) were assessed in detail and an variable importance analysis was conducted. For the variable importance analysis, a modified permutation method was applied. All models were created in accordance with the OECD principles for QSARs (OECD, 2004). Fig. 1 provides an overview of the methods applied here, which are also described in more detail down below.

2.2. Data collection

For the collection of dissolution data, the ecotoxicity dataset extracted by Balraadjsing et al. (Balraadjsing et al., 2024) was used as a starting point and was subsequently modified by including or excluding observations. Information regarding the dataset in Balraadjsing et al. (Balraadjsing et al., 2024) (*e.g.* literature search, collected variables) can be found in the supplementary materials of that paper. Briefly, the data collected in Balraadjsing et al. (Balraadjsing et al., 2024) included features regarding the physico-chemical parameters of MNPs, their exposure conditions and toxic effects towards species. Additionally, molecular descriptors were calculated using the Alvadesc calculator (www.alvascience.com/alvadesc) through the OCHEM platform (www.ochem.eu).

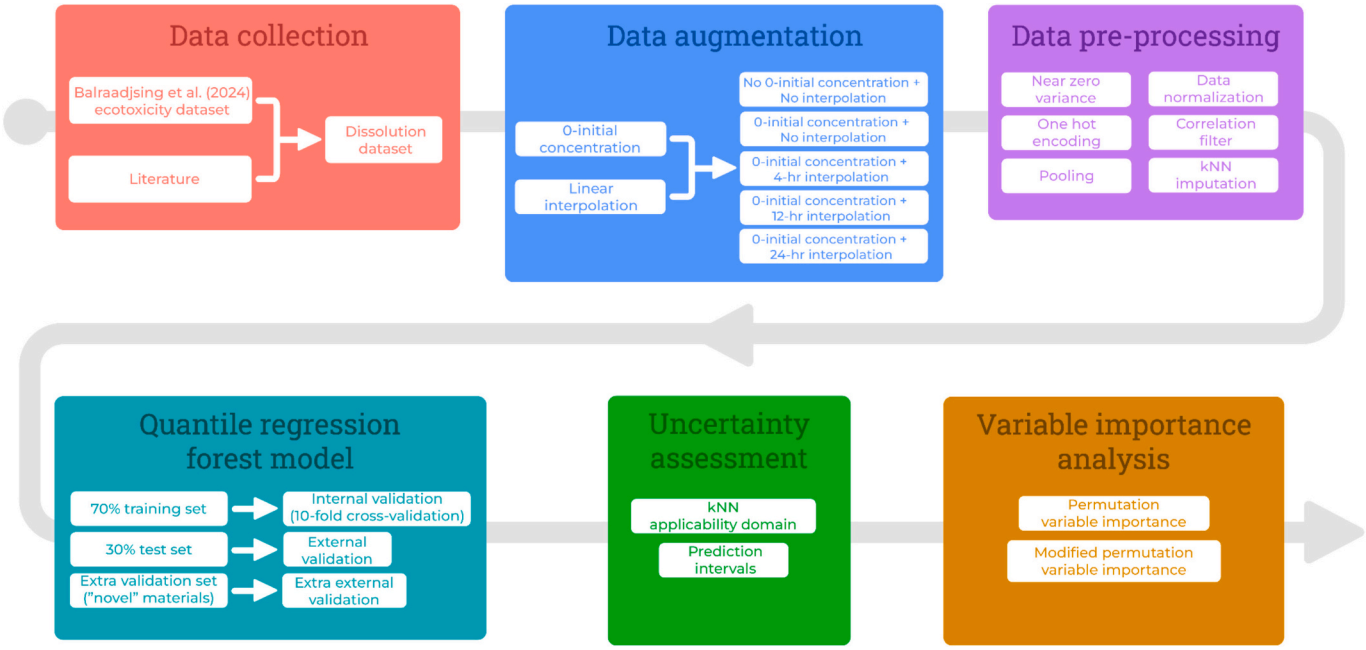


Fig. 1. Schematic overview of data collection and modelling process.

From the initial dataset, papers with no reported or measured dissolution data were firstly excluded. Furthermore, only features related to the physico-chemical parameters of MNPs and experimental conditions were retained (features related to species and toxicity were not deemed relevant here). Following this, dissolution data was extracted from the remaining papers for different time points and initial MNP concentrations. During this process, the original ecotoxicity dataset was modified where necessary to match the conditions of the dissolution experiments (*e.g.* due to different experimental conditions used than the toxicity experiments). This means that the database was expanded depending on the study, due to the particle dissolution being measured at multiple time points during their experiments. The extracted dissolution related features included exposure time (h), the initial MNP concentration (μg/L), concentration of released ions (μg/L) and the percentage of dissolution (defined as the ratio of released metal ions concentration to the initial MNPs concentration (μg/L)). The latter was chosen as the model's outcome. Further information regarding the collected variables can be found in Supplementary material S1.1.

Special attention was given to ensuring the quality and reliability of the data, therefore data points with incomplete dissolution data or inconsistent information (*e.g.* unclear how experiments were conducted or ambiguous experimental conditions) were excluded. Data cleaning and filling in of gaps were done as in Balraadjsing et al. (Balraadjsing et al., 2024). Data for the top materials with the most observations were initially selected for creating QSPRs and the remaining data for materials with very few observations were discarded from modelling but used for extra external validation (the purpose of this dataset is explained further below). The materials that were used for modelling included: Ag, $BaTiO_3$, CeO_2, Cu, CuO, NiO, Se, SiO_2, TiO_2 and ZnO. While data on Fe_3O_4, Fe_2O_3, Al_2O_3, Ni, Co_3O_4 MNPs were also relatively abundant, these materials were excluded from modelling due to poor performance as also described in Balraadjsing et al. (Balraadjsing et al., 2024). These materials were added to the extra external validation dataset.

2.3. Data augmentation

Data augmentation in ML is the process of artificially increasing the size and diversity of a dataset by synthetically creating new instances to build a larger dataset for model training (Maharana et al., 2022). The goal here was to investigate whether model performance would improve by artificially increasing the dataset. Data augmentation was done here in two approaches: 1) duplicating the dataset and replacing the dissolution related parameters (the initial particle concentration and ion release) with zeros (hereafter referred to as 0-initial concentration), and 2) the use of linear interpolation at different time points.

Theoretically, at 0-initial particle concentration there cannot be any ion release (there can be no dissolution if there are no MNPs present in the solution). This allowed us to duplicate the entire dataset and replace the initial concentration and ion release with zeros.

Furthermore, when the extent of dissolution of two time points is known then the ion release at any time point in between should be within the range of those points. For example, if at 1 h the ion release is 10 % and at 3 h it is 20 %, then the ion release at 2 h should be between in between 10 and 20 %. To achieve this, we applied linear interpolation to generate synthetic data for papers where dissolution data was reported at different time points. This was done for papers that had dissolution data for three time points or more (for the same material and conditions) to prevent the creation of highly unreliable data. Linear interpolation was applied every 4 h, 12 h or 24 h to find out what effect data augmentation had on the performance of the models. The amount of synthetically created data increased with decreasing time points chosen for interpolation (interpolation every 4 h has the most augmented data, while interpolation every 24 h has the least).

In total five datasets were generated: original dataset (no interpolation + no 0-initial concentration), only 0-initial concentration (no interpolation), 4-h interpolation +0-initial concentration, 12-h interpolation +0-initial concentration, and 24-h interpolation +0-initial concentration.

2.4. Data pre-processing

An essential part of machine learning is the pre-processing of data to make it more suitable for ML algorithms. Data was pre-processed as described in Balraadjsing et al. (Balraadjsing et al., 2024). Briefly, this included the implementation of a near-zero variance filter, excluding some variables from the filter and handling categorical variables by converting them to dummy variables using one-hot encoding. Infrequent levels of categorical variables were pooled into an "other" category. The data was normalized and a correlation filter was applied, removing variables with a Pearson correlation coefficient above 0.9. Missing data was filled in where possible according to Balraadjsing et al. (Balraadjsing et al., 2024) and otherwise imputed using k-Nearest Neighbors (kNN) imputation.

2.5. Machine learning-based Model Development

Following pre-processing, regression QSPR models were developed based on a ML approach, with the outcome being the amount of ions released (in %). While it can be beneficial to generate models based on different ML algorithms, we created QSPRs based on only one algorithm here, quantile regression forests (a variation of the random forest algorithm). Due to the nature of the target variable, the outcome of it is limited to a certain range (0 to 100 % ion release). An issue with various ML algorithms is that they cannot be easily constrained to produce predictions within such a range. However, random forests (and therefore also quantile regression forests) fail at extrapolating beyond the ranges of the training set (Takoutsing and Heuvelink, 2022), which can be seen as a "feature not a bug" here. This "flaw" will automatically constrain the outcome between 0 and 100 % and prevent the models from producing predictions beyond that range. Furthermore, the quantile regression forests algorithm was favored here over the random forest algorithm for its ability to generate prediction intervals (Sun et al., 2022).

The created dataset was randomly split into a training and test set (70 % train and 30 % test) using stratified sampling. Stratified sampling prevents certain materials from dominating the training or test set. Following this, QSPR models were trained for each of the previously mentioned five datasets. 10-fold cross-validation was used for internal validation, to assess robustness, to perform hyperparameter optimization, and to prevent overfitting (Arlot and Celisse, 2010).

The goodness-of-fit, robustness and predictivity of the developed QSPR models were evaluated using various performance metrics. For the regression models, the assessment included the concordance correlation coefficient (CCC), Huber loss, mean absolute error (MAE), root mean squared error (RMSE), and R^2. The R^2 and CCC range from 0 to 1 where higher scores indicate a better fit. In contrast, the Huber loss, MAE, and RMSE are context-dependent metrics (based on the outcome, which in this case is the ion release), with lower values indicating better performance. Good and similar results from both internal- and external validation indicate that models are robust and not overfitted (Rendón et al., 2011).

External validation was done on the test set and on the discarded data (that included data for materials not used for modelling). Extra external validation, using the discarded data, was conducted to gain insight into how the models perform when applied to "novel" or "unseen" materials. All models were created in accordance with the OECD principles for predictive models (European Commission Environment Directorate General, 2014).

2.6. Model uncertainty (applicability domain and prediction intervals)

The applicability domain (AD) defines the ranges within which the

QSPR model can provide reliable predictions (Bunmahotama et al., 2022). The AD in this work was generated by using the k-nearest neighbor approach on Euclidean distances (structural domain) of a training dataset to define the space where accurate predictions can be made with a specified level of 95 % and 99 % thresholds confidence as described in Balraadjsing et al. (Balraadjsing et al., 2022). Observations within the 95 % confidence interval are considered reliable. Data points between the 95 % and 99 % confidence intervals should be treated with caution. Data outside the 99 % confidence interval are considered unreliable extrapolations because they differ significantly from the training data. To complement the applicability domain, prediction intervals were also generated, offering more insight into prediction uncertainty. By design, quantile regression forests inherently produce these prediction intervals.

2.7. Variable importance analysis

To identify the key factors influencing the rate of ion release during dissolution, we conducted a variable importance analysis based on a permutation method. This involved permuting variables and ranking them based on their impact on the model's predictive performance. The process works by taking the training data and shuffling one variable while keeping the rest constant. The models are then re-run, and the performance metrics are re-calculated and compared to the baseline model performance. Increase or decrease of the model errors are indicative of the variable being important. Variables with higher importance scores were deemed more influential in determining the dissolution behavior of MNPs.

However, a potential downside of using permutation variable importance for assessing variable importance is the key assumption of variable independence. Due to the fact that the method shuffles one variable at a time while keeping everything else constant, it poses a problem for features that are interconnected with each other (*e.g.* the surface area increases as the particle size decreases). These relationships get lost, leading to over- or underestimated importance scores. To circumvent this issue, we modified the method by nesting features together (that are known to be related to each other) prior to permuting. The nested variables were then shuffled together, preserving this connection and the importance scores were then calculated as previously stated. Variables that were nested together included: crystallinity_anatase + crystallinity_rutile, shape + primary_diameter + length + surface_area, salinity + water_hardness + conductivity.

3. Results

3.1. Dataset screening and composition

After selecting and screening 98 published papers from the initial dataset by Balraadjsing et al. (Balraadjsing et al., 2024), a total of 1648 dissolution datapoints were obtained based on 21 different MNPs including 17 metal oxide nanoparticles and 4 metallic nanoparticles (Fig. 2). In the dataset, Ag MNPs made up most of the observations (47 %). Data on Ag, ZnO and CuO MNPs together, constituted more than 80 % of the whole dataset, while the other materials were investigated to a lesser extent.

3.2. Model performance

Model performance was evaluated through internal- (training set) and external (test set) validation based on different performance metrics (Table 1). Results showed that the model without any data augmentation (no interpolation and no 0-initial concentration) had a reasonably good fit ($R^2 = 0.749$; CCC = 0.859) but relatively high errors compared to the other models (RMSE = 9.55; MAE = 3.58). Introducing 0-initial concentration without interpolation resulted in a reduced fit ($R^2 = 0.690$; CCC = 0.818) and slightly higher RMSE (9.74) but a significantly lower MAE (2.63). Interpolating data with 24 h and 0-initial concentration reduced errors (RMSE = 7.33; MAE = 1.78) and improved the fit,

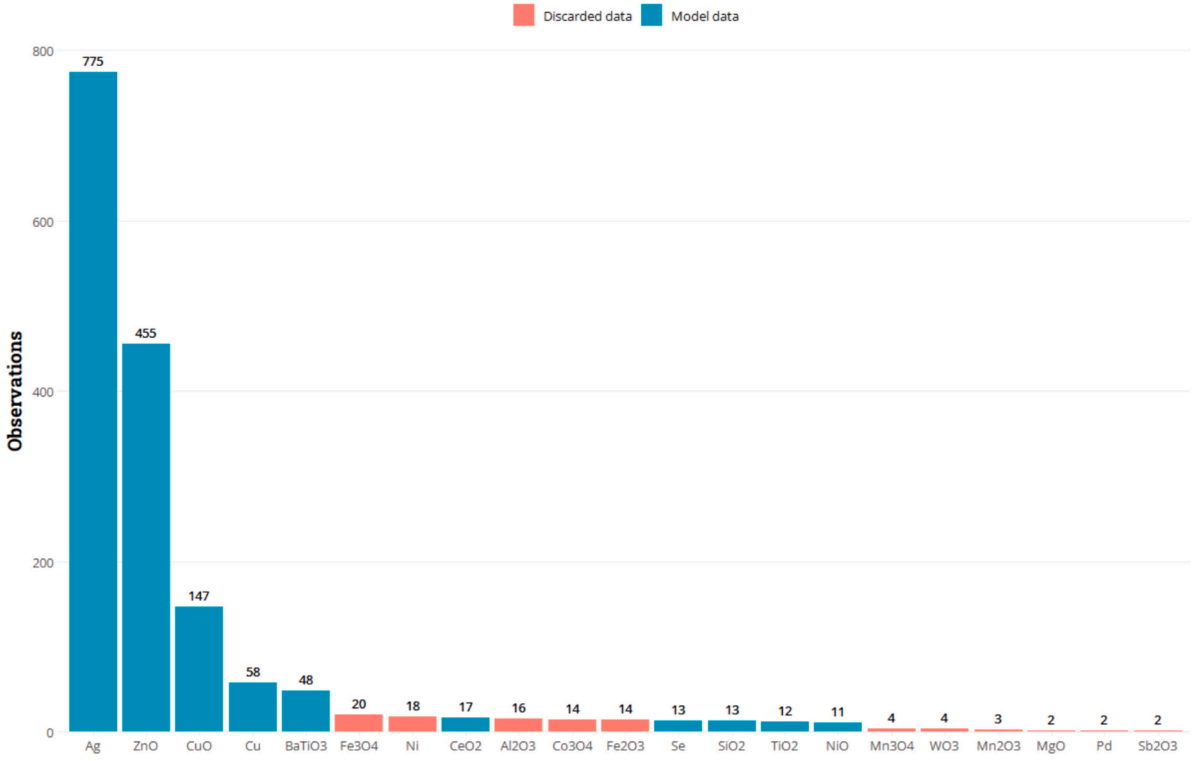


Fig. 2. Amount of observations present in the dataset for each of the MNPs after data collection from the literature. Blue bars = data for the materials were used for modelling, red bars = data that were excluded from modelling and used for extra external validation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Model performance metrics based on the whole dataset after external validation (on the test set) (Int. = interpolation, h = hours, Zeros = 0-initial concentration, RMSE = Root Mean Squared Error, MAE = Mean Absolute Error, R^2 = R squared, CCC = Concordance Correlation Coefficient).

Int.	Zeros	Train size	Test size	RMSE	MAE	R^2	CCC	Huber loss
No	No	1083	466	9.55	3.58	0.749	0.859	3.27
No	Yes	1735	746	9.74	2.63	0.690	0.818	2.44
24 h	Yes	2217	951	7.33	1.78	0.777	0.880	1.62
12 h	Yes	3016	1295	7.47	1.76	0.781	0.883	1.62
4 h	Yes	6268	2687	4.67	0.841	0.903	0.949	0.730

($R^2 = 0.777$; CCC = 0.880) relative to the model without any augmented data. Continuing this trend, interpolation with 12 h and 0-initial concentration maintained similar improvements with relatively similar errors (RMSE = 7.47; MAE = 1.76) and an improved fit ($R^2 = 0.781$; CCC = 0.883) relative to the previous model. The most notable improvement was seen with 4-h interpolation plus 0-initial concentration, (RMSE = 4.67; MAE = 0.841; $R^2 = 0.903$; CCC = 0.949), demonstrating superior model performance relative to the model without any data augmentation and other models with less data augmentation applied. Based on the performance metrics alone, this would indicate that when the amount of data augmentation increases, model performance improves significantly. Finally, comparison to internal validation results (Supplementary material S2.1) showed similar performance between models and suggests no signs of model overfitting. When models overfit, they tend to perform well on data they have seen before (training set) but poorly when they see new data (test set). Therefore, by comparing internal- and external validation results between each other, signs of over- or underfitting can be detected. With the results showing relatively similar performance results between the training- and test sets, it can be concluded that the models neither overfitted nor underfitted.

Furthermore, this pattern of improvement is also highlighted in the actual *versus* predicted plots for the created model (Figs. 3a, 3b and Supplementary material S3). Only the best performing model (Fig. 3b) and model with no data augmentation (Fig. 3a) are shown here, the remaining actual *versus* predicted plots for other models can be found in Supplementary material S3. The results showed that while the majority of the data points are centered around the diagonal line, these points were largely made up of the augmented data (points labeled "zeros" and "interpolate" in Fig. 3b and Supplementary material S3). Contrary to the performance metrics (Table 1), visual inspection of Figs. 3a and 3b indicated that synthetically adding more data may not necessarily improve the model's original fit or its ability to predict the original data (non-augmented data), and that the improvement in model performance may be misleading. This can also be seen when evaluating model performance based on only the original (non-augmented) data (Supplementary material S2.2). Here the augmented data was removed during calculations to gain insight into how the augmented data influenced the calculated performance metrics. When the augmented data was not part of the performance evaluation, models performed relatively similar or worse compared to the model with no augmented data. In other words, the models did not necessarily get better at predicting the outcome when data was augmented. Instead, the augmented data influenced the final calculated performance metrics. Performance metrics calculated for each material group are provided in Supplementary material Tables S3.1a and S3.1b, but will not be discussed further in this paper due to issues described in the supplementary materials. Fig. 3a: Actual *versus* predicted plots for ion release of MNPs after evaluation on the test set (model without data augmentation). X- axis represents actual ion release percentage (%) from experimental data, Y- axis represents predicted ion release percentage (%) from model. Data points are color coded "original" (original non-augmented data).

3.3. Model uncertainty (applicability Domain and prediction intervals)

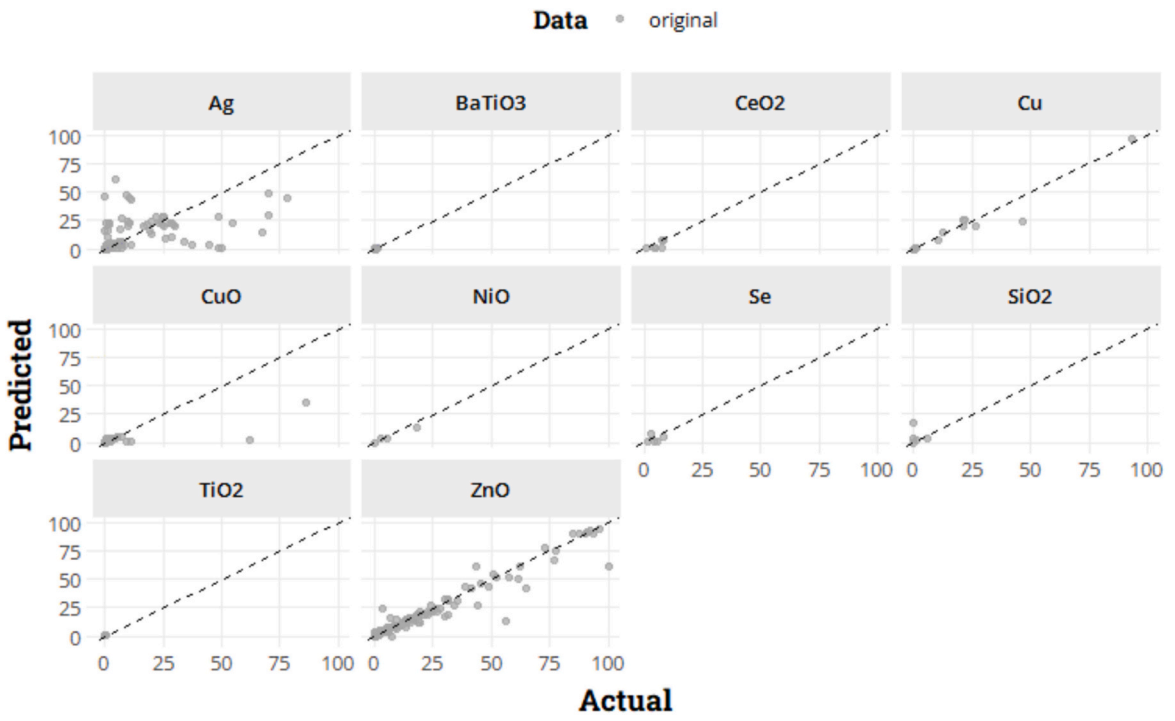
To assess the limitations of the models created here, its domain of applicability was investigated through a kNN approach. Results for only the best performing model are mentioned here (remaining results can be found in Supplementary material S4). The AD results in this study revealed that a substantial proportion (2606 out of 2687) data points are categorized as "reliable" (within the 95 % limit) (Supplementary material Table S4.1). This indicated that the training dataset effectively covered the feature space of the test set, thus enhancing the model's reliability for these predictions. However, 17 data points were classified as "cautious" (between the 95 % and 99 % limits (near the borders of the AD)) and 64 were "unreliable" (outside the 99 % limit (outside AD)). This classification suggested potential areas where the model may exhibit reduced predictive performance, highlighting the limitations in the current training dataset and indicating regions that required additional data or model refinement.

When comparing the AD results in combination with the actual *versus* predicted plots (Fig. 4), it was shown that datapoints labeled "cautious" and "unreliable" did not necessarily have predictions with higher errors. On the contrary, the majority of the points highlighted as "cautious" and "unreliable" were all relatively close to the diagonal line. This highlighted how datapoints could be considered outside of the assessed AD but still had fairly accurate predictions, and *vice versa*, datapoints inside the AD could have predictions with large errors. Fig. 4 also showed how datapoints related to $BaTiO_3$, CeO_2, Se, SiO_2, TiO_2 MNPs and to a lesser extent Cu were frequently considered to be around the borders of, or outside of the AD.

By employing the quantile regression forests algorithm, we successfully generated prediction intervals, which complements the results of the applicability domain and helps gain more insight in to the uncertainty of the predictions. Results for only the best performing model (4-h interpolation +0-initial concentration) is presented here to illustrate the value of complementing the AD with prediction intervals. Fig. 5 shows the majority of the generated prediction intervals intersecting with the diagonal line, indicating that the "true" values were within the generated prediction intervals. Some data points displayed large prediction intervals (Ag, ZnO and to a lesser extent Cu MNPs), indicating high uncertainty and could be deemed less reliable predictions.

3.4. Variables importance

Permutation variable importance results for the best performing regression model (4-h interpolation +0-initial concentration) generated crucial insights into the factors driving the dissolution process. The nested variable importance methodology, which group related variables together during permutation, showed that the initial concentration of MNPs played an important role in the dissolution process, consistently ranking as the most influential variable (Fig. 6). Besides the initial concentration, the following exposure conditions were generally influential: temperature, water hardness, salinity, conductivity, exposure time and pH. The key physico-chemical properties included: surface area, shape, primary diameter, length and coating. The surface area, shape and primary diameter were grouped together in the nested results, and they contributed equally to the dissolution process (Fig. 6). Although most molecular descriptors displayed lower importance, three descriptors were ranked reasonably high in both permutation analyses: mean atomic Sanderson electronegativity (Me), mean atomic van der



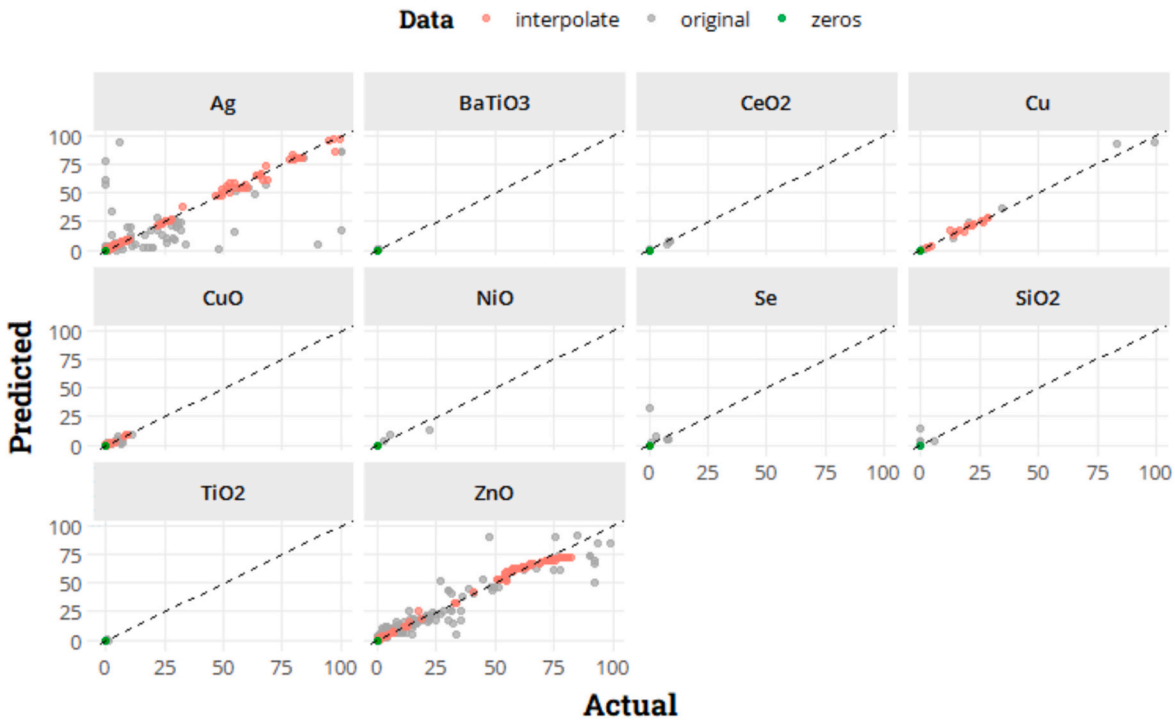


Fig. 3b. Actual *versus* predicted plots for ion release of MNPs after evaluation on the test set (model based on 4-h interpolation and 0-initial concentration). X- axis represents actual ion release percentage (%) from experimental data, Y- axis represents predicted ion release percentage (%) from model. Data points are color coded "interpolate" (linear interpolated data), "original" (original non-augmented data) or "zeros" (0-initial concentration augmented data).

Waals volume (Mv) and sum of atomic Sanderson electronegativities (Se). Due to the modifications applied to the variable importance method here and permutation being applied prior to pre-processing, the importance of all features was calculated. This included variables that were supposed to be filtered but they automatically get zero importance as a result of pre-processing being applied by the models later on in the process.

For comparison, the variable importance based on its original method (without nesting variables together) was also calculated (Supplementary material S5). These results also showed several variables related to the exposure conditions and physico-chemical properties ranked relatively high in variable importance scores, whereas the majority of molecular descriptors scored low (Supplementary material S5). However, the ranking of variables differed relative to the nested results,

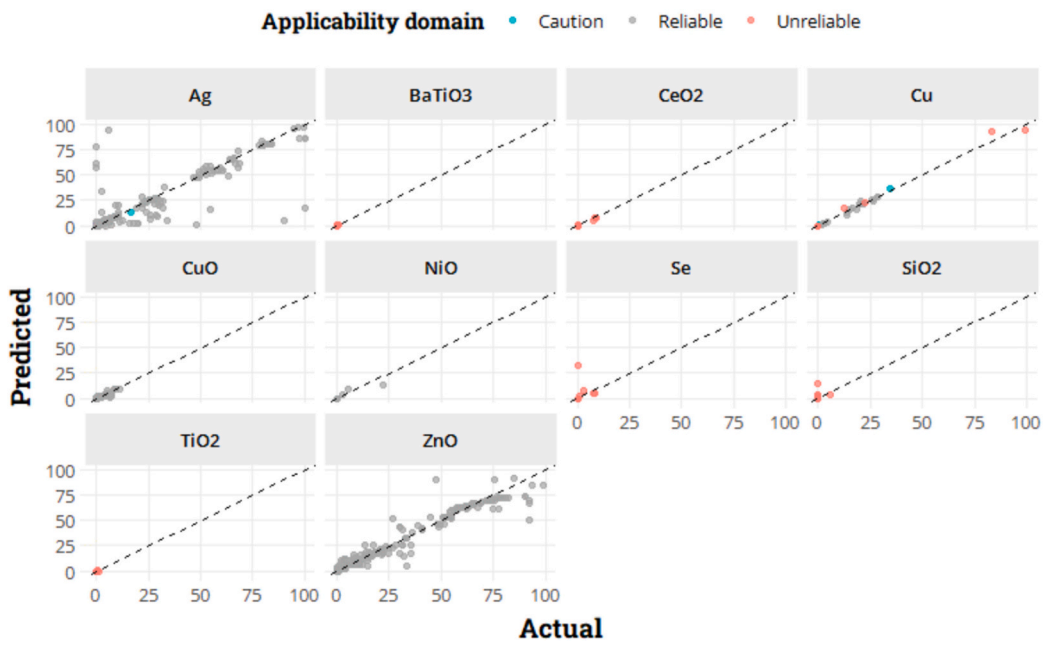


Fig. 4. Actual *versus* predicted plot for the quantile regression forests after evaluation on the test set (4-h interpolation +0-initial concentration). X- axis represents actual ion release percentage (%) from experimental data, Y- axis represents predicted ion release percentage (%) from model. Data points are color coded "caution", "reliable" or "unreliable" based on the results of the definition of the applicability domain.

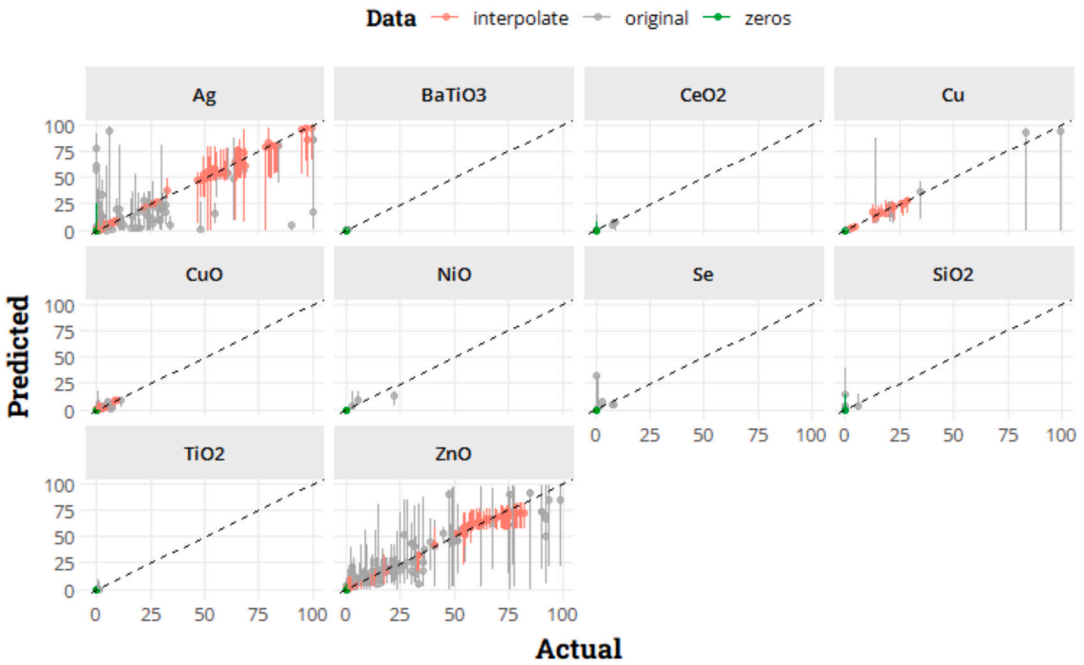


Fig. 5. Actual *versus* predicted plot of the best performing model (4-h interpolation +0-initial concentration) divided per MNP. X- axis represents actual ion release percentage (%) from experimental data, Y- axis represents predicted ion release percentage (%) from model. Data points are color coded "interpolate" (linear interpolated data), "original" (original non-augmented data) or "zeros" (0-initial concentration augmented data). The vertical bars represent the prediction intervals.

mainly relating to the variables that were nested together. While the surface area, shape, length and primary diameter scored high when nested together, they are relatively of lesser importance when evaluated individually. This also applies to other variables such as water hardness and salinity. Comparing the results of both methods indicates how the importance of variables can be underestimated when they share relationships with other variables (and are therefore not independent variables). By nesting such dependent variables together, their importance is assessed together instead of evaluating their importance individually (as independent variables). This helps in uncovering variables that may not be important on its own (as an independent variable) but is important because it is connected to others.

3.5. *Performance on novel materials*

After assessment of model performance on the extra validation dataset (data discarded for modelling), it was highlighted that the best performing model (4 h interpolation +0-initial concentration) did not

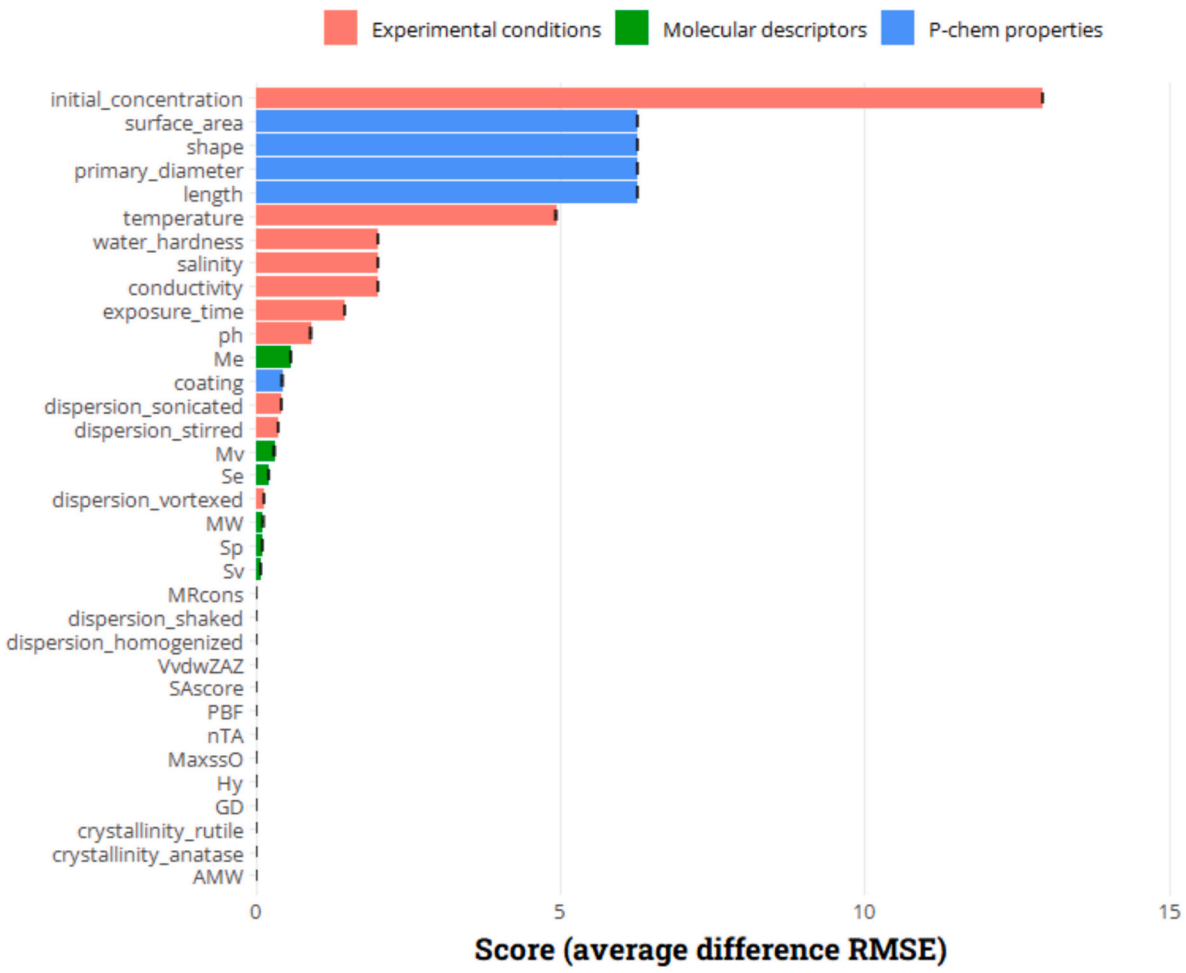


Fig. 6. Variable importance scores after modification by nesting variables together (average of the RMSE difference with the base model).Horizontal black bars represent the standard errors and the colored bars represent the different variable groups: "experimental conditions" (red), "molecular descriptors" (green) and "physical-chemical properties" (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

produce completely accurate results (Fig. 7). Furthermore, when also calculating the applicability domain for the discarded dataset, this resulted in the majority of the data points being considered unreliable with the exception of datapoints for Pd and Ni MNPs, which were labeled "caution". Six data points were flagged as "reliable". These results are in accordance with the results shown based on the test set (Figs. 4 and 5) where the AD results do not always line up with the accuracy of predictions.

4. Discussion

In this study, we developed QSPR models that predict the percentage dissolution of MNPs at any time point and under different experimental conditions as can be found in laboratory test settings. Moreover, the performance and robustness of the created QSPR models were investigated while also examining variable importance and the effect of data augmentation. The created models showed overall good performance despite the relatively small dataset used in terms of data volume. Model performance increased with the amount of synthetically added data. These QSPR models are of importance because they can aid in filling in data gaps regarding particle dissolution and identifying the factors that mediate MNP dissolution.

4.1. Model performance and effect of data augmentation

The regression QSPR models were generated based on a quantile regression forest algorithm and all models displayed relatively high predictivity and robustness that can be deemed good performing models ($R^2 \geq 0.6$) (Klimenko et al., 2016; Roy et al., 2019). The benefit of using quantile regression forests over other algorithms was due to its ability to produce predictions within a specific range (0–100 % in this case) while also having the ability to produce prediction intervals, as stated in the materials and methods section. While other algorithms may potentially perform better or have better fits, improbable predictions can be produced (< 0 % or > 100 %). It is possible that neural networks, using a sigmoid activation function could potentially address the need for predictions within a bounded range (Na et al., 2022), but further experimentation and tuning would be necessary to determine its effectiveness in this context. Although sigmoid activation functions are typically applied to classification neural networks, it may also work in this case due to the nature of the outcome variable (percentage dissolution) here. This is a topic for further study.

The robustness of a model should be evaluated through both internal- and external validation, both of which were presented as a series of performance metrics in this paper (Table 1 and Supplementary material S2.1). Among the different degrees of augmentation applied, the 4-h interpolation with 0-initial concentration model emerged as the best performing model. This approach demonstrated the lowest RMSE compared to other models, indicating superior performance and accuracy. The results from Table 1 highlighted the impact data augmentation can have on model performance. The most effective strategy involved interpolation with smallest intervals (4 h) combined with adding 0-

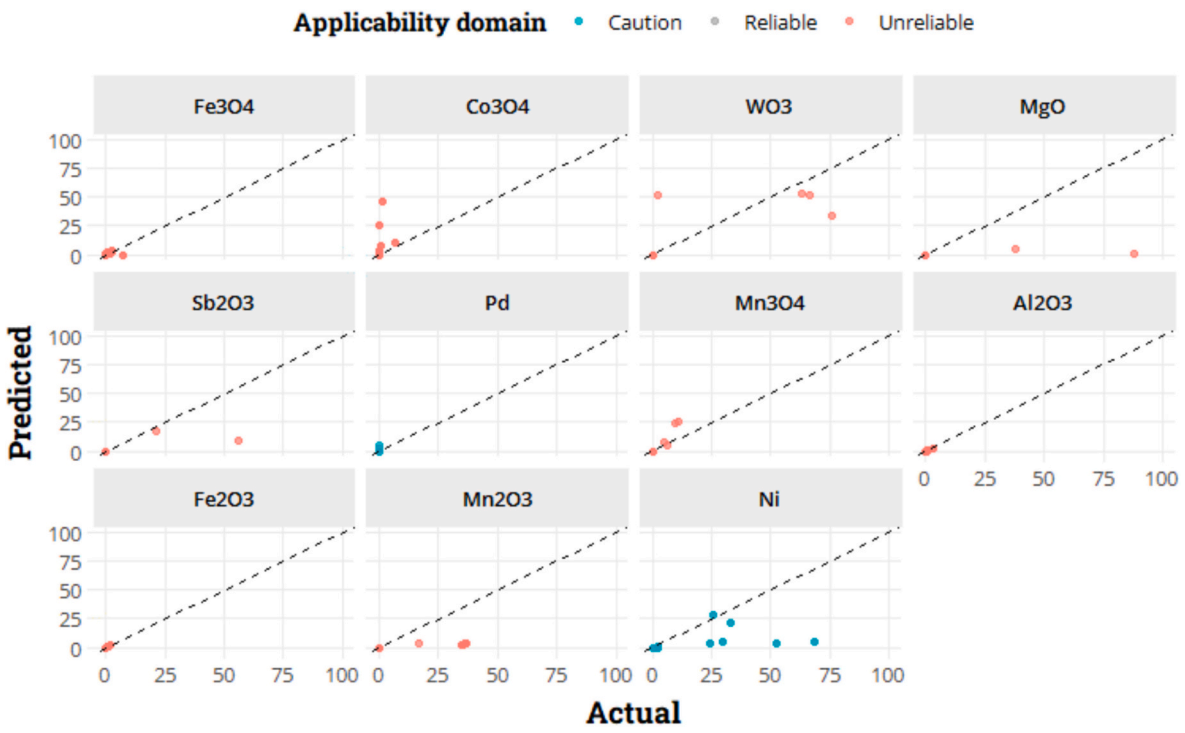


Fig. 7. Actual *versus* predicted plot of the best performing model (4-h interpolation +0-initial concentration) on extra validation dataset, divided per MNP. X- axis represents actual ion release percentage (%) from experimental data, Y- axis represents predicted ion release percentage (%) from model. Data points are color coded "caution", "reliable" or "unreliable" based on the results of the definition of the applicability domain (Reliable 6; Caution 26; Unreliable 157).

initial concentration, which consistently improved the predictive accuracy across all metrics (Mikolajczyk et al., 2015; Toropova et al., 2011). This approach would suggest that the augmented data provided the model with a denser and more informative dataset, allowing it to better capture the dissolution process dynamics. In another study, data augmentation was also applied to artificially increase their dataset and overcome the limited available data for modelling (Furxhi et al., 2023). Although different methods were applied, similar results were found where performance improved significantly after data augmentation.

However, a deeper look at the model performance results after different degrees of data augmentation provided more insight into the effects of data augmentation. Interestingly, the model performance was generally similar when being evaluated without the augmented data (Supplementary material S2.2), suggesting that the augmented data did not improve the algorithms' predictive abilities or its ability to learn from the original (non-augmented) training data. When looking at the actual *versus* predicted plots (Fig. 3b and Supplementary material S3), it can also be seen that the majority of the augmented data were centered around the diagonal line while this was not the case for the non-augmented data (labeled "original"). In other words, while synthetically increasing the dataset may improve performance when looking at the model as a whole, it did not necessarily improve the model's fit towards the original non-augmented data. The algorithms seemed to perform exceptionally well on the augmented data (perhaps overfitting) and this may overestimate model performance to some extent. More research is needed into the topic of data augmentation, the influence it has on QSPR (or QSAR) model outcomes and how to properly assess its effects on model performance.

4.2. Model Uncertainty (applicability domain and prediction intervals)

The majority of test data fell within the thresholds set by the calculated applicability domain (Fig. 4), thus most predictions made using the test set can be considered reliable. Regarding model performance on different groups, the applicability domain was highly similar across all groups. Results highlighted how datapoints related to $BaTiO_3$, CeO_2, Se, SiO_2, TiO_2 and to a lesser extent Cu were frequently considered to be around the borders of, or outside of the AD. This suggested that the models may not be suitable for such materials. The idea of the applicability domain is to assess the limits of the chemical space so that the model is not used to extrapolate beyond it. Extrapolating beyond the chemical space of the training data can increase the errors and uncertainty of predictions (Basei et al., 2019; Sahigara et al., 2012). However, results obtained here showed that the results of the applicability domain did not line up well with the accuracy of predictions (*e.g.* a point being outside the assessed chemical space can still have an accurate prediction and *vice versa*). It is possible that the applicability domain used here overestimates data points and the thresholds should be set more conservatively for more reliability (*e.g.* a lower k value or instead of a 95–99 % limit, use an 80–85 % limit). The value for k and the thresholds set by the kNN applicability domain methodology are done so arbitrarily (Furxhi et al., 2020), with k usually being the square root of the amount of observations and a 95th percentile threshold based on the calculated distances. By setting the limits more conservatively, the applicability domain will get smaller and while this is means that a model can be applied to a limit amount of chemical/materials, it may have increased reliability. For example, (Sahigara et al., 2013) suggested the use of a k value of $n^{1/3}$ (n being the amount of observations), which will lead to a more conservative applicability domain than when the square root is used. The topic of applicability domains of predictive models using MNP datasets should be investigated further in addition to how their thresholds affect its reliability.

A downside of the applicability domain is that it only gives insight into whether a data point is within or outside of the assessed chemical space (it is qualitative in nature). The applicability domain does not say anything about the quantitative uncertainty of the predictions. To address this, the use of prediction intervals can provide more insight into prediction uncertainty. When complemented with the applicability domain, prediction intervals can help gain more insight into the reliability of predictive models. By quantifying the uncertainty associated

with each prediction, researchers and practitioners can gauge the confidence level of the model's outputs. Results of the generated prediction intervals (Fig. 5) showed the majority of the intervals intersecting with the diagonal line (the "true" value), indicating that the calculated intervals captured the true values within them. Relatively larger intervals were seen for Ag and ZnO MNP data, indicating larger uncertainty. This is in contrast with the applicability domain results where instead $BaTiO_3$, CeO_2, Se, SiO_2, TiO_2 and to a lesser extent Cu were generally considered to have larger uncertainty as previously stated. However, the prediction interval results here (Fig. 5) line up better with the accuracy of the predictions compared to the applicability domain: the data points with the largest errors belonged to Ag and ZnO data. Furthermore, it should be noted that prediction intervals do not always capture the "true" value within its intervals and therefore are not a better methodology to assess the uncertainty of predictive models (Balraadjsing et al., 2024). We recommend them to be used together with the applicability domain to gain a better picture of whether predictions can be trusted or not.

4.3. Variable importance analysis

Understanding how the properties and variables in the models created here interact with each other is critical for identifying the dissolution mechanisms of MNPs (Hedberg et al., 2019). Variable importance was evaluated using permutation tests, which underscored the significance of features pertaining to both aquatic conditions and physico-chemical properties of MNPs (Figs. 6 and Supplementary material S5). The comprehensive approach of this study, combining both nested and unnested variable importance analyses, provides a holistic understanding of the factors driving MNP dissolution.

The results demonstrate significant advancements in understanding the dissolution dynamics of MNPs and underscore the critical factors that influence this process. The identification of the initial concentration of MNPs as the most influential variable aligns with findings from other studies. For instance, research by (Avramescu et al., 2023) highlighted that the initial concentration of silver nanoparticles significantly affected their dissolution rate in aquatic environments: lowering the initial concentration (10 mg L^{-1} *versus* 100 mg L^{-1}) significantly increased the relative solubility (% of total concentration) of nano-ZnO and nano-MnO_2 in both water and Dulbecco's Modified Eagle's Medium (DMEM), which corroborated our results that higher initial concentrations lead to a less pronounced dissolution process. Furthermore, the findings of our study on the importance of surface area, shape and primary diameter, which contributed equally to the dissolution and ranked as second the most important factors are also consistent with previous research (Zhao et al., 2017). It is generally accepted that a reduction in nanoparticle size generally results in an increased dissolution rate due to the increased surface area (Roduner, 2006). Meanwhile, a decreased MNP size is associated with an increase of the surface area, which in turn promotes dissolution by providing a greater number of surface sites that can participate in the dissolution process (Borm et al., 2006). Similarly, as reported in previous literature (Garnier et al., 2019), when controlling the shape of the nanoparticles, the MNPs with a well-defined surface structure could be obtained. Therefore, these three factors are interconnected and may interact with each other. A nested approach could be used as an appropriate method to help preserve the interactions among them. Nesting variables together during a permutation variable importance analysis, can preserve variable interactions and give better insight into crucial variables that mediate the dissolution process. When using the traditional (unnested) permutation approach, the importance of certain variables may be underestimated despite being interconnected to other crucial variables.

Furthermore, comparing the nested and unnested results, the unnested approach provided additional insights into specific factors such as conductivity and temperature. Conductivity is related to the ionic strength of the medium. Higher ionic strength can shield the electrostatic repulsion between charged nanoparticles, potentially increasing aggregation and reducing surface area available for dissolution (Oh et al., 2012). Similarly, the role of temperature in enhancing dissolution rates has been well-documented, as seen in the study of Seyed et al. (Majedi et al., 2014), who demonstrated that higher temperatures increased the dissolution of ZnO MNPs by reducing the surface charge in aqueous solutions.

Regarding the variable importance results, it should be noted that the unbalanced structure of the data set may have also affected the results. In Zhou's study (Zhou et al., 2022), TiO_2 MNPs and Ag MNPs, which were both sensitive to the light conditions, comprised more than half of the data. Consequently, the importance of illumination might be overestimated, indicating the need for a more balanced dataset to make the results more representative. It is noticeable that particle coating is not considered as very important in explaining the extent of dissolution (Figs. 6 and Supplementary material S5), given that it is known that the dissolution happens on the surface of the particles and that surface chemistry significantly influences ion release (Liu and Hurt, 2010). The reason can be attributed to the quality of the data set used here: details about the coating were only categorized as "coated" or "uncoated" due to the large amount of data and variation in coatings used, causing the model to perhaps underestimate the importance of specific coating details. A better approach for dealing with the particle coating, such as classifying the different coatings by their functions or calculating molecular descriptors associated with them, can help aid in uncovering their importance in better detail. This should be investigated further.

4.4. Model performance on novel materials

Although numerous QSPR models are developed with outstanding performance metrics that meet the criteria for a good predictive model, they frequently face criticism for their lack of practical accuracy when applied to novel materials (Balraadjsing et al., 2024; Basei et al., 2019; Yu et al., 2023). Therefore, we also investigated model performance towards "novel" materials not included in the training set (discarded dataset). Through applying models on these MNPs (Fe_3O_4, Co_3O_4, WO_3, MgO, Sb_2O_3, Pd, Mn_3O_4, Al_2O_3, Fe_2O_3, MgO Mn_2O_3 and Ni), it was revealed that models performed poorly (Fig. 7), regardless of whether observations were within or outside of the applicability domain. The model generally underestimated the ion release of most data points. It could be argued that the "rare" nanoparticles in the discarded data should be regarded as dissimilar to those in the training- and test sets and this can also be seen to some extent in Fig. 7. However, it can also be seen that the outcome of the AD did not line up well with the accuracy of the predictions. For Ni data points, they were all considered "reliable" but the accuracy of the predictions were rather poor. On the other hand, Al_2O_3, Fe_2O_3 and Fe_3O_4 were considered "unreliable" extrapolations according to the AD, but the model produced rather accurate results. These type of predictions have also been identified in (Balraadjsing et al., 2024) and underscores the importance of evaluating models's outcome carefully beyond the training- and test sets as also argued in said paper. This is important as QSPR and QSAR models are created with the intention of being applied towards "novel" chemicals or materials but they are rarely evaluated beyond their traditional training- and test sets (which usually have similar substances between the both of them). It should be noted that the scarcity of data makes it challenging to perform such an evaluation. As shown here, the created QSPR model can perform poorly, underestimating the ion release of materials not included in the training set, regardless of whether the applicability domain considers them within the chemical space of the model or not. Therefore, we recommend the models to not be used beyond the materials included in the training set, even if the AD considers them to be within the assessed chemical space.

4.5. Future work and perspectives

The predictive dissolution models presented here, integrated particle and medium characteristics to enhance its applicability across diverse scenarios. The data upon which the models were built came from laboratory experiments, published in literature. Most of the available experimental data in literature have been executed at relatively high initial particle concentrations, which has also been previously reported for ecotoxicity experiments (Holden et al., 2016). This results into the fact that the percentage ions released is controlled by the initial concentration of nanomaterials in the system as also shown by the QSPR model variable importance results. This is obvious, as high nanomaterial concentrations can exceed the equilibrium dissolution levels of specific nanomaterials, but this may not be environmentally relevant, as existing studies have shown that environmental concentrations of nanomaterials are quite low (Keller and Lazareva, 2014; Lead et al., 2018). At such lower concentrations, kinetic dissolution may be more relevant rather than percentage released ions. However, it should be noted that the research on dissolution kinetics is in its infancy and such parameters are rarely reported in current MNP dissolution literature, making it difficult to implement into QSPR models.

When MNPs are released into the aquatic environment, they undergo various transformation processes and are exposed to other elements. This means that the dissolution process is driven by multiple different factors simultaneously, some of which were not considered in this study. For instance, the dissolution on the surface of certain types of MNPs such as Ag particles, can easily change due to sulfidation or redox transformations (Kang and Park, 2021; Lead et al., 2018). However, the use of such information within QSPRs is severely limited as they are frequently not reported or easily comparable between studies.

The models developed here also help gain insight into the dissolution process and into QSPR model building. Additionally, the QSPR models can also aid in filling in data gaps. QSPR models for nanomaterials rarely focus on dissolution as the outcome despite being considered one of the more crucial transformation processes a nanomaterial undergoes due to its role in nanotoxicity towards species. The creation of such predictive dissolution models can be challenging due to the lack of available data. Experimental dissolution data can be very scarce, inconsistent and incomplete. This is primarily due to the lack of standardized methodologies and poor reporting. This resulted in the developed model being built on a large amount of missing data which had to be filled in by hand or through imputation. This obviously has consequences for the *in silico* models and thus we iteratively plea for better data reporting practices.

Although the models created here displayed good performance, our results also highlighted pitfalls when assessing QSPR models. Namely, applicability domain results may not always give a clear indication regarding the reliability of predictions. Future work could focus on the ins and outs of applicability domains and how they are influenced by the data and how different input parameters affect their outcomes as previously stated in the discussion. The kNN applicability domain is frequently used in QSPR or QSAR modelling, usually with standard albeit arbitrary methods to calculate input parameters. We speculate that it is possible that for smaller and noisy datasets (such as the one used here), a more conservative approach to the input parameters might yield better results. This would result in a smaller applicability domain and more "extreme" thresholds but perhaps increase the reliability of results. However, this must be investigated in depth.

With the scarcity of available data for modelling purposes, data augmentation can be an attractive option to help artificially increase datasets for machine learning purposes. However, this is relatively new and thus poorly explored with regards to QSPR and QSAR modelling. As shown and argued throughout this paper, data augmentation should be used with caution and its effect on models must be rigorously assessed. Other methods of data augmentation should be explored in future studies and how they can be safely applied for QSAR/QSPR modelling. More research is needed into the topic of data augmentation, the influence it has on QSPR (or QSAR) model outcomes and how to properly assess its effects on model performance.

5. Conclusion

By constructing a comprehensive dissolution database from published literature and employing ML techniques, we developed QSPR models capable of predicting the ion release of MNPs. A ML based-QSPR approach might find practical applications in designing new and safer nanoparticles (safe-by-design). Our findings emphasized the critical role that the initial particle concentration, size related MNP properties and the experimental conditions play in the dissolution process. Furthermore, we also highlight the need for caution and proper model evaluation when applying data augmentation. While data augmentation can help overcome the small nanomaterial datasets, it is crucial that its effects are assessed in depth, otherwise misleading conclusions could be garnered and model performance may be overestimated. Furthermore, the benefits of generating prediction intervals were also highlighted for a better understanding of prediction uncertainty and to complement the applicability domain. Prediction intervals are quantitative and thus give more insight into the uncertainty of predictions relative to the (qualitative) applicability domain. Applicability domains may also not always provide reliable results into whether predictions can be trusted or not and complementing it with the prediction intervals can aid here.

These models add to the body of growing QSAR/QSPR modelling literature for nanomaterials and provide insight into how such models can be built, how to assess them in depth and aid in filling data gaps. The results and QSPR models presented here can serve to support researchers in designing dissolution experiments by allowing for pre-screening of concentration ranges prior to conducting experiments. Likewise, the models also facilitate the setting of optimal time points for dissolution quantification, streamlining research efforts and thereby saving valuable time and resources. Ultimately, the approach presented in this study has the potential to generate full dissolution curves by producing ion release predictions at different time points and estimating the time needed to achieve a steady-state concentration. Finally, we plea for better data reporting practices when experiments are conducted and published, so that data could be reused more easily for modelling purposes and prevent unnecessary gaps within data sets.

Notes

The authors declare no competing financial interest.

CRediT authorship contribution statement

Yuchao Song: Writing – original draft, Investigation, Funding acquisition, Data curation, Conceptualization. **Surendra Balraadjsing:** Writing – review & editing, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **Willie J.G.M. Peijnenburg:** Writing – review & editing, Supervision. **Martina G. Vijver:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the ERC-C grant entitled EcoWizard no. 101002123 (granted to MGV). The Chinese Scholarship Council (CSC) is gratefully acknowledged for its financial support to YS (201906320061).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.impact.2025.100547.

Data availability

Data will be made available on request.

References

Abbas, Q., Yousaf, B., Amina, Ali, Munir, M.A.M., El-Naggar, A., Rinklebe, J., Naushad, M., 2020. Transformation pathways and fate of engineered nanoparticles (ENPs) in distinct interactive environmental compartments: A review. Environ. Int. 138 (October 2019), 105646. https://doi.org/10.1016/j.envint.2020.105646.

Amde, M., Liu, J., Tan, Z.Q., Bekana, D., 2017. Transformation and bioavailability of metal oxide nanoparticles in aquatic and terrestrial environments. A review. Environ. Pollut. 230, 250–267. https://doi.org/10.1016/j.envpol.2017.06.064.

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Statistics Surveys 4, 40–79. https://doi.org/10.1214/09-SS054.

Avramescu, M.L., Chénier, M., Beauchemin, S., Rasmussen, P., 2023. Dissolution behaviour of metal-oxide nanomaterials in various biological media. Nanomaterials 13 (1). https://doi.org/10.3390/nano13010026.

Balraadjsing, S., Peijnenburg, W.J.G.M., Vijver, M.G., 2022. Exploring the potential of in silico machine learning tools for the prediction of acute Daphnia magna nanotoxicity. Chemosphere 307 (P2), 135930. https://doi.org/10.1016/j.chemosphere.2022.135930.

Balraadjsing, S., J.G.M. Peijnenburg, & Vijver, M. G. (2024). Building species trait-specific nano-QSARs: model stacking, navigating model uncertainties and limitations, and the effect of dataset size. Environ. Int., 188(May), 108764. doi:https://doi.org/10.1016/j.envint.2024.108764.

Basei, G., Hristozov, D., Lamon, L., Zabeo, A., Jeliazkova, N., Tsiliki, G., Marcomini, A., Torsello, A., 2019. Making use of available and emerging data to predict the hazards of engineered nanomaterials by means of in silico tools: A critical review. NanoImpact 13 (2018), 76–99. https://doi.org/10.1016/j.impact.2019.01.003.

Borm, P., Klaessig, F.C., Landry, T.D., Moudgil, B., Pauluhn, J., Thomas, K., Trottier, R., Wood, S., 2006. Research strategies for safety evaluation of nanomaterials, part V: role of dissolution in biological fate and effects of nanoscale particles. Toxicol. Sci. 90 (1), 23–32. https://doi.org/10.1093/toxsci/kfj084.

Brown, K.A., Brittman, S., Maccaferri, N., Jariwala, D., Celano, U., 2020. Machine learning in nanoscience: big data at small scales. Nano Lett. 20 (1), 2–10. https://doi.org/10.1021/acs.nanolett.9b04090.

Bunmahotama, W., Vijver, M.G., Peijnenburg, W., 2022. Development of a Quasi–Quantitative Structure–Activity Relationship Model for Prediction of the Immobilization Response of *Daphnia magna* Exposed to Metal-Based Nanomaterials. Environ. Toxicol. Chem. 41 (6), 1439–1450. https://doi.org/10.1002/etc.5322.

Ding, X., Pu, Y., Tang, M., Zhang, T., 2022. Environmental and health effects of graphene-family nanomaterials: potential release pathways, transformation, environmental fate and health risks. Nano Today 42, 101379. https://doi.org/10.1016/j.nantod.2022.101379.

European Commission Environment Directorate General, 2014. Guidance Document on the Validation of (Quantitative)Structure-Activity Relationships [(Q)Sar] Models. February, 1–154.

Furxhi, I., Murphy, F., Mullins, M., Arvanitis, A., Poland, C.A., 2020. Practices and trends of machine learning application in nanotoxicology. Nanomaterials 10 (1), 1–32. https://doi.org/10.3390/nano10010116.

Furxhi, I., Kalapus, M., Costa, A., Puzyn, T., 2023. Artificial augmented dataset for the enhancement of nano-QSARs models. A methodology based on topological projections. Nanotoxicology 17 (6–7), 529–544. https://doi.org/10.1080/17435390.2023.2268163.

Gardea-Torresdey, J.L., Rico, C.M., White, J.C., 2014. Trophic transfer, transformation, and impact of engineered nanomaterials in terrestrial environments. Environ. Sci. Technol. 48 (5), 2526–2540. https://doi.org/10.1021/es4050665.

Garnier, E., Vidal-Iglesias, F.J., Feliu, J.M., Solla-Gullón, J., 2019. Surface structure characterization of shape and size controlled Pd nanoparticles by cu UPD: A quantitative approach. Front. Chem. 7 (July), 1–11. https://doi.org/10.3389/fchem.2019.00527.

Gupta, S., Basant, N., 2018. Predictive modeling: solubility of C60 and C70 fullerenes in diverse solvents. Chemosphere 201, 361–369. https://doi.org/10.1016/j.chemosphere.2018.02.174.

Hedberg, J., Blomberg, E., Odnevall Wallinder, I., 2019. In the search for Nanospecific effects of dissolution of metallic nanoparticles at freshwater-like conditions: A critical review. In. Environ. Sci. Technol. 53 (8), 4030–4044. https://doi.org/10.1021/acs.est.8b05012.

Holden, P.A., Gardea-Torresdey, J.L., Klaessig, F., Turco, R.F., Mortimer, M., Hund-Rinke, K., Cohen Hubal, E.A., Avery, D., Barceló, D., Behra, R., Cohen, Y., Deydier-Stephan, L., Ferguson, P.L., Fernandes, T.F., Herr Harthorn, B., Henderson, W.M., Hoke, R.A., Hristozov, D., Johnston, J.M., et al., 2016. Considerations of environmentally relevant test conditions for improved evaluation of ecological hazards of engineered nanomaterials. Environ. Sci. Technol. 50 (12), 6124–6145 doi: 10/gm2xzt.

Huang, Y., Li, X., Cao, J., Wei, X., Li, Y., Wang, Z., Cai, X., Li, R., Chen, J., 2022. Use of dissociation degree in lysosomes to predict metal oxide nanoparticle toxicity in immune cells: machine learning boosts nano-safety assessment. Environ. Int. 164 (August 2021), 107258. https://doi.org/10.1016/j.envint.2022.107258.

Kang, J.S., Park, J.W., 2021. Silver ion release accelerated in the gastrovascular cavity of Hydra vulgaris increases the toxicity of silver sulfide nanoparticles. Environ. Toxicol. Chem. 40 (6), 1662–1672. https://doi.org/10.1002/etc.5017.

Keller, A.A., Lazareva, A., 2014. Predicted releases of engineered nanomaterials: from global to regional to local. Environ. Sci. Technol. Lett. 1 (1), 65–70. https://doi.org/10.1021/ez400106t.

Klimenko, K., Kuz'min, V., Ognichenko, L., Gorb, L., Shukla, M., Vinas, N., Perkins, E., Polishchuk, P., Artemenko, A., Leszczynski, J., 2016. Novel enhanced applications of QSPR models: temperature dependence of aqueous solubility. J. Comput. Chem. 2045–2051. https://doi.org/10.1002/jcc.24424.

Lead, J.R., Batley, G.E., Alvarez, P.J.J., Croteau, M.-N., Handy, R.D., McLaughlin, M.J., Judy, J.D., Schirmer, K., 2018. Nanomaterials in the environment: behavior, fate, bioavailability, and effects-an updated review: nanomaterials in the environment. Environ. Toxicol. Chem. 37 (8), 2029–2063. https://doi.org/10.1002/etc.4147.

Li, X., Lenhart, J.J., Walker, H.W., 2010. Dissolution-accompanied aggregation kinetics of silver nanoparticles. Langmuir 26 (22), 16690–16698. https://doi.org/10.1021/la101768n.

Li, J., Wang, C., Yue, L., Chen, F., Cao, X., Wang, Z., 2022. Nano-QSAR modeling for predicting the cytotoxicity of metallic and metal oxide nanoparticles: A review. Ecotoxicol. Environ. Saf. 243 (April), 113955. https://doi.org/10.1016/j.ecoenv.2022.113955.

Liu, J., Hurt, R.H., 2010. Ion release kinetics and particle persistence in aqueous nano-silver colloids. Environ. Sci. Technol. 44 (6), 2169–2175. https://doi.org/10.1021/es9035557.

Lowry, G.V., Avellan, A., Gilbertson, L.M., 2019. Opportunities and challenges for nanotechnology in the Agri-tech revolution. Nat. Nanotechnol. 14 (6), 517–522. https://doi.org/10.1038/s41565-019-0461-7.

Loza, K., Diendorf, J., Sengstock, C., Ruiz-Gonzalez, L., Gonzalez-Calbet, J.M., Vallet-Regi, M., Köller, M., Epple, M., 2014. The dissolution and biological effects of silver nanoparticles in biological media. J. Mater. Chem. B 2 (12), 1634–1643. https://doi.org/10.1039/c3tb21569e.

Maharana, K., Mondal, S., Nemade, B., 2022. A review: data pre-processing and data augmentation techniques. Global Transitions Proc. 3 (1), 91–99. https://doi.org/10.1016/j.gltp.2022.04.020.

Majedi, S.M., Kelly, B.C., Lee, H.K., 2014. Combined effects of water temperature and chemistry on the environmental fate and behavior of nanosized zinc oxide. Sci. Total Environ. 496, 585–593. https://doi.org/10.1016/j.scitotenv.2014.07.082.

Mikolajczyk, A., Gajewicz, A., Rasulev, B., Schaeublin, N., Maurer-Gardner, E., Hussain, S., Leszczynski, J., Puzyn, T., 2015. Zeta potential for metal oxide nanoparticles: A predictive model developed by a nano-quantitative structure-property relationship approach. Chem. Mater. 27 (7), 2400–2407. https://doi.org/10.1021/cm504406a.

Na, M., Nam, S.H., Moon, K., Kim, J., 2022. Development of a nano-QSAR model for predicting the toxicity of nano-metal oxide mixtures to Aliivibrio fischeri. Environ. Sci. Nano 10 (1), 325–337. https://doi.org/10.1039/d2en00672c.

Odzak, N., Kistler, D., Behra, R., Sigg, L., 2015. Dissolution of metal and metal oxide nanoparticles under natural freshwater conditions. Environ. Chem. 12 (2), 138–148. https://doi.org/10.1071/EN14049.

OECD, 2004. OECD Environment Health and Safety Publications Series on Testing and Assessment No. 49—Report from the Expert Group On (Quatitative) Structure-Activity Relationships [(Q)SARs] On the Principles For the Validation of (Q)SARs. Organisation for Economic Co-Operation and Development (OECD).

Oh, M.H., Kim, N.S., Kang, S.M., 2012. Effect of conductivity of the aqueous solution on the size of printable nanoparticle. J. Nanotechnol. 2012. https://doi.org/10.1155/2012/467812.

Poynton, H.C., Chen, C., Alexander, S.L., Major, K.M., Blalock, B.J., Unrine, J.M., 2019. Enhanced toxicity of environmentally transformed ZnO nanoparticles relative to Zn ions in the epibenthic amphipod Hyalella azteca. Environ. Sci. Nano 6 (1), 325–340. https://doi.org/10.1039/C8EN00755A.

Rendón, E., Abundez, I.M., Gutierrez, C., Zagal, S.D., Arizmendi, A., Quiroz, E.M., Elsa Arzate, H., 2011. A comparison of internal and external cluster validation indexes. In: Applications of Mathematics and Computer Engineering - American Conference on Applied Mathematics, AMERICAN-MATH'11, 5th WSEAS International Conference on Computer Engineering and Applications, CEA'11, pp. 158–163.

Roduner, E., 2006. Size matters: why nanomaterials are different. Chem. Soc. Rev. 35 (7), 583–592. https://doi.org/10.1039/b502142c.

Roy, J., Ghosh, S., Ojha, P.K., Roy, K., 2019. Predictive quantitative structure-property relationship (QSPR) modeling for adsorption of organic pollutants by carbon nanotubes (CNTs). Environ. Sci. Nano 6 (1), 224–247. https://doi.org/10.1039/c8en01059e.

Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., Todeschini, R., 2012. Comparison of different approaches to define the applicability domain of QSAR models. Molecules 17 (5), 4791–4810. https://doi.org/10.3390/molecules17054791.

Sahigara, F., Ballabio, D., Todeschini, R., Consonni, V., 2013. Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. J. Chemother. 5 (5), 1–9. https://doi.org/10.1186/1758-2946-5-27.

Sun, Y., Huang, Y., Yang, M., 2022. Ultra-short-term wind power interval prediction based on fluctuating process partitioning and quantile regression Forest. Front. Energy Res. 10 (April), 1–13. https://doi.org/10.3389/fenrg.2022.867719.

Takoutsing, B., Heuvelink, G.B.M., 2022. Comparing the prediction performance, uncertainty quantification and extrapolation potential of regression kriging and random forest while accounting for soil measurement errors. Geoderma 428 (May), 116192. https://doi.org/10.1016/j.geoderma.2022.116192.

Tangaa, S.R., Selck, H., Winther-Nielsen, M., Khan, F.R., 2016. Trophic transfer of metal-based nanoparticles in aquatic environments: A review and recommendations for future research focus. Environ. Sci. Nano 3 (5), 966–981. https://doi.org/10.1039/c5en00280j.

Toropov, A.A., Sizochenko, N., Toropova, A.P., Leszczynski, J., 2018. Towards the development of global nano-quantitative structure–property relationship models: zeta potentials of metal oxide nanoparticles. Nanomaterials 8 (4). https://doi.org/10.3390/nano8040243.

Toropova, A.P., Toropov, A.A., Benfenati, E., Gini, G., Leszczynska, D., Leszczynski, J., 2011. CORAL: QSPR models for solubility of [C60] and [C70] fullerene derivatives. Mol. Divers. 15 (1), 249–256. https://doi.org/10.1007/s11030-010-9245-6.

Tripathy, N., Hong, T.K., Ha, K.T., Jeong, H.S., Hahn, Y.B., 2014. Effect of ZnO nanoparticles aggregation on the toxicity in RAW 264.7 murine macrophage. J. Hazard. Mater. 270, 110–117. https://doi.org/10.1016/j.jhazmat.2014.01.043.

Tschiche, H.R., Bierkandt, F.S., Creutzenberg, O., Fessard, V., Franz, R., Giese, B., Greiner, R., Haas, K.H., Haase, A., Hartwig, A., Hund-Rinke, K., Iden, P., Kromer, C., Loeschner, K., Mutz, D., Rakow, A., Rasmussen, K., Rauscher, H., Richter, H., et al., 2022. Environmental considerations and current status of grouping and regulation of engineered nanomaterials. Environ. Nanotechnol. Monit. Manag. 18 (May), 100707. https://doi.org/10.1016/j.enmm.2022.100707.

Yu, M., Shi, Y., Liu, X., Jia, Q., Wang, Q., Luo, Z.H., Yan, F., Zhou, Y.N., 2023. Quantitative structure-property relationship (QSPR) framework assists in rapid mining of highly thermostable polyimides. Chem. Eng. J. 465 (April), 142768. https://doi.org/10.1016/j.cej.2023.142768.

Zhai, Y., Hunting, E.R., Wouterse, M., Peijnenburg, W.J.G.M., Vijver, M.G., 2017. Importance of exposure dynamics of metal-based nano-ZnO, -cu and -Pb governing the metabolic potential of soil bacterial communities. Ecotoxicol. Environ. Saf. 145 (July), 349–358. https://doi.org/10.1016/j.ecoenv.2017.07.031.

Zhao, Y., Wang, Y., Ran, F., Cui, Y., Liu, C., Zhao, Q., Gao, Y., Wang, D., Wang, S., 2017. A comparison between sphere and rod nanoparticles regarding their in vivo biological behavior and pharmacokinetics. Sci. Rep. 7 (1), 1–11. https://doi.org/10.1038/s41598-017-03834-2.

Zhou, Y., Wang, Y., Peijnenburg, W., Vijver, M.G., Balraadjsing, S., Fan, W., 2022. Using machine learning to predict adverse effects of metallic nanomaterials to various aquatic organisms. Environ. Sci. Technol. https://doi.org/10.1021/acs.est.2c07039.